DEEP SPACE COMMUNICATION AND NAVIGATION STUDY

Volume 2--Communication Technology

May 1, 1968
FINAL REPORT
Contract No. NAS 5—10293
Contracting Officer: R. M. Keefe, NASA
Technical Monitor: J. E. Miller, NASA
Project Manager: J. S. Cook, BTL

Prepared by Bell Telephone Laboratories, Incorporated
Whippany Road, Whippany, New Jersey
On behalf of Western Electric Company, Incorporated
83 Maiden Lane, New York, New York
For National Aeronautics and Space Administration
Goddard Space Flight Center
Greenbelt, Maryland

ABSTRACT

This study provides a comparison of alternative means for high data rate communication (about 10^6 b/s) from deep space probes, and indicates the extent to which orbiting spacecraft can aid deep space navigation. Emphasis is on the communication problem. A special effort has been made to delineate practical and theoretical constraints on communication from a distance of 1 to 10 AU at microwave, millimeter, and optical frequencies (1 to 100 GHz and 20 to 0.2 microns wavelength), and to indicate promising avenues for extending the art.

The interrelationship between fundamental theory, device characteristics, and system performance has received particular attention in this study. Specific missions have been synthesized, and problems of visibility, Doppler variation, handover, acquisition tracking, and synchronization have been investigated in order to discover the limitations imposed by practical system considerations.

This study was initiated and directed by Ira Jacobs.

CONTENTS
VOLUME 2

Chapter 1. Microwave Systems

Chapter 2. Millimeter Systems

Chapter 3. Optical Technology

Chapter 4. Optical Communications

New Technology. Pulse Position Modulation For Optical Communication

ILLUSTRATIONS
VOLUME 2

TABLES

VOLUME 2

CHAPTER I. MICROWAVE SYSTEMS

The performance of present deep-space communication systems must be improved to meet the needs of future missions. Methods for obtaining this improvement are being considered at different regimes of frequency. The capabiiities of a microwave communication system, one of the important contenders, are examined in this section. Knowledge of the ground receiving system, the propagation medium, and the spacecraft transmitting system are required to determine the performance characteristics.

The Deep Space Instrumentation Facility (DSIF) is considered first. The DSIF has provided ground system support for space programs in the past and presumably will in the future. The investment in these facilities is a factor in the choice of operating frequencies for the future deep space probes. The past, present, and projected deep space receiving system capabilities of this facility are listed.

The transmitter capabilities of future spacecraft are considered next. Device limitations are discussed, and the power, weight, efficiency, and lifetime of future spacecraft transmitters for the 2 to 16 GHz band are listed.

The spacecraft antenna is an important part of the transmitter system. Five different types of spacecraft antennas are considered. Available gain is obtained as a function of weight and frequency for antennas from 1 to 100 GHz. Spacecraft antenna pointing requirements are also considered.

The effects of the communication medium can be determined by a study of atmospheric effects – attenuation, sky noise, and refraction under clear weather, snow, and ice, and cloudy and rainy conditions. Except in the case of communication through rain and clouds, where there is inadequate knowledge of path characteristics, the effects are generally well known and small. The limitations are discussed and a method is recommended which will enable the lack of knowledge to be overcome.

The choice of microwave or millimeter operating frequency is strongly influenced by the cost and performance of large ground station antennas, and this subject is discussed. Mathematical models relating diameter, cost, gain, frequency, and rms surface tolerance are developed for antennas with and without a radome. Any three variables can be determined if the other two are provided. Examples of the use of these models are given.

Finally, the communication performance of biorthogonal modulation systems is considered. Available trade-offs and advantages of the system are considered for a single digital transmission link from the deep space probe to the Earth.

1. DSIF

The Deep Space Instrumentation Facility (DSIF) is an element of the NASA Deep Space Network (DSN) and is managed by the Jet Propulsion Laboratory. The DSIF at present includes stations at Goldstone (4), Australia (2), South Africa, Spain, Cape Kennedy, and the Ascension Islands. It provides, as a minimum, a three-longitude network for deep space communication in support of the NASA unmanned space exploration program. In addition, it permits the "cautious integration" of reliable, operational field devices and laboratory techniques without jeopardizing current spacecraft programs. The Goldstone stations – Pioneer, Echo, Venus, and Mars (named after the projects for which they were first used) -- are the testing grounds for most of the technological advances.

Since 1958 the DSIF has evolved with steady improvement in receiving and transmitting capabilities. The DSIF performance (past, present, and future) is listed in Table 11. There have been considerable variations in the reported system performance (in particular in the noise temperature) because of development activities at one or more stations which extend the state of the art but are not used until a later date in an actual operational system.

Of particular note in the increased performance over the years are: (1) the increased antenna diameter from 85 to 210 feet, (2) the advent of the maser which is primarily responsible for decreasing the system's noise temperature from 1450°K, and (3) the increase in transmitter power from 10 to 400 kW.

Some of the lesser known JPL efforts offer promise for the future. The X-band (8.448 GHz) work has developed to

Table 11

DSIF PERFORMANCE

	1960	1962	1965	1967	1970
Antenna					
Noise temperature (°K)	–	–	–	10	7-10
Diameter (ft)	85	85	85	85	210
Efficiency	55	58	47	70	77
Transmitter					
Frequency (MHz)	378	960	2400	2400	2400
Power (kW)	10	13	100	100	400
Receiver, noise temperature (°K)	–	–	10	7	7
System, noise temperature (°K)	1450	65	65	55	25

such an extent that it has been used as a backup in the DSN and was actually used to track Mariner IV for three days. The receiver, incorporated at the Venus station in the 85-foot antenna, uses a maser with an 18°K equivalent noise temperature. System noise temperature is 38°K. It is to be used at the Mars station to provide information about the performance of the 210-foot antenna at this frequency.

JPL has proposed three design goals for advanced antenna systems spanning the next decade:[1]

1. Increase the transmitter power-handling capability of the feed and associated components to a minimum of 500 kW CW with a design goal of 2000 kW CW.
2. Increase the antenna gain by extending present feedhorn techniques to the transmitter frequency band in addition to the receiver band.
3. Increase the antenna figure of merit (the ratio of aperture efficiency to antenna noise temperature)[2] over the receiver band.

Initial measurements indicate the design goal of obtaining 2000 kW components may be achievable.[3]

2. SPACE TRANSMITTERS

Space transmitters at microwave frequencies, as considered in this section, include oscillators and amplifiers at frequencies from 2 to 16 GHz. Although a general review of the subject is presented, little space is given to material covered in detail in other literature. The section covers only devices and techniques appropriate for deep space communication. Therefore, low-power devices (less than 10 watts, e.g., negative grid vacuum tubes) and multikilowatt devices (>10 kW) are excluded.

Microwave power sources in general are either vacuum tubes or solid-state devices. Vacuum-tube sources include crossed-field and linear beam devices. (Negative grid tubes are not competitive in power or in maximum operating frequency.) Solid state sources include transistor, tunnel diode, varactor, Gunn, IMPATT, and LSA devices.

Various aspects of microwave power output devices have been studied previously. These studies provide an accurate state of the art description and in addition provide valuable sources for bibliographic material. Two survey articles[4,5] discuss microwave tubes in broad terms with trends and capabilities highlighted. DeLoach[6] presents a survey of capabilities and limitations of solid state devices.

Several articles have discussed trends of space transmitters. Feldman[7] discusses the relative characteristics of various satellite output devices and provides an excellent bibliography. Others have specifically discussed the use of traveling wave tubes for spacecraft transmitters.[9, 10] Specific advantages and disadvantages for space use of known types of microwave power sources are also listed in a recent report.[11]

For the devices considered most likely to be adaptable for use as high-power space transmitters, power, efficiency, weight, and lifetime are discussed while bandwidth and gain are not discussed in detail. The (3 dB) bandwidths of the devices all exceed 10 MHz (which is adequate for the advanced modulation techniques discussed in other chapters of this report). Gains are adequate and typically are greater than 30 dB. The actual gain required involves a tradeoff for each specific system between the weights, efficiencies, and power output of the driver stage and amplifier.

Problems associated with the transmitter such as power supplies, prime power (Appendix 1), and thermal difficulties are discussed in this chapter. Limitations imposed by these problems are considered after the devices themselves are discussed.

2.1 Vacuum Tubes

Crossed-field and linear beam devices have many subclasses, which differ mainly in the form of the rf circuit that interacts with the electron beam. Of the crossed-field devices, the amplitron and the CFA (crossed-field amplifier) especially deserve consideration.

The amplitron is a continuous cathode, re-entrant beam, backward-wave amplifier. The efficiency of the amplitron (50 to 80 percent) exceeds that of other microwave tubes. In addition, the ratio of power output to weight is comparatively high. There has been considerable interest in the amplitron. A Raytheon amplitron (QKS 1300) is to be used on the LEM (Lunar Excursion Module).[7] (Primary communications, however, are provided by a TWT chain in the orbiting command module). The tube provides 25 watts of output power in S-band and 16 dB gain with an overall transmitter efficiency (including the power supply) of 35 to 40 percent.

However, several basic limitations exist which make the amplitron unattractive for long-life, deep space probes. More than one amplifier might be necessary for high power requirements because the tubes have a limited gain (approximately 10 to 17 dB). The lifetime of the amplitron is also questionable. Life tests on the QKS 1300 have demonstrated only a 3000 to 4000 hour lifetime.[8] The life is limited by the lack of cathode emission due to back-bombarding electrons – an inherent feature of the amplitron. An extension of the lifetimes presently obtained, if possible, for higher power amplitrons, will require a significantly different cathode design. No such effort currently exists.[8]

The CFA is an injected-beam, forward-wave amplifier. This tube overcomes the objections listed for the amplitron – the gain is higher (45 dB), and lifetime is no longer severely limited by cathode bombardment because the device has a collector. However, its efficiency is less (35 to 50 percent). The resulting tube has no significant advantage over the more commonly used linear-beam, traveling-wave tube. In fact, one of the difficulties is the lack of interest in the CFA; it is expected that problems will not be adequately explored and future development will be impaired.

Linear beam tubes include the traveling-wave tube (an injected-beam, forward-wave amplifier) and the klystron (an injected-beam, standing-wave amplifier). Both are proved microwave devices and both are being considered for use in active deep space programs.[12,13]

The traveling-wave tube is about seven years younger than the klystron. Its wideband capability brought it to prominence as an important advance. The wide bandwidth is due to the characteristics of the helix interaction circuit. Bandwidth is limited to about an octave by the couplers at the input and output of the helix. Power is limited to 1 to 2 kW or slightly greater. The maximum beam voltage is of the order of 10 kV because at higher voltages the desired mode of propagation begins to deteriorate and the rf interaction with the beam is reduced. Current is limited by the electron optics.

A variation of the helix structure (commonly called the ring-bar or counterwound helix) increases the average power-handling capability slightly. The coupled-cavity circuit provides still higher average power. It has a more massive structure and is easier to cool. In general, the cost of this high-power capability is decreased bandwidth (5 percent overall bandwidth is typical). Also, a solenoid is required for focusing; thus weight and size are increased, overall efficiency is decreased because of power supplied to the solenoid, and power supply requirements are increased.

Figure 19 is a map, on a coordinate system of power vs frequency, which displays a boundary for some types of TWT's. Points for helix and ring bar TWT's lie in the cross-hatched region. Points representing the highest known power obtained at 3 GHz[14] and 16 GHz with helix type TWT's are plotted. The line connecting these two points has a slope of about -2. It has been shown theoretically that this slope should exist when scaling traveling wave tubes[15] (or, as frequently stated, Pf^2 is constant) and that the line is determined by the thermal design technology. A line is also drawn at a power of 2 kW since this is approximately the maximum power for helix and ring bar TWT's.

All presently space-qualified TWT's lie in the cross-hatched region. Emphasis has been placed on the efficiency and reliability of these lighter, lower power, helix-type satellite tubes. Since existing techniques are adequate for present requirements, there has been no great endeavor to push for the higher powers. However, helix and ring bar TWT's may ultimately be extended to higher frequencies. To do so will require techniques (which do not now exist) to overcome the thermal problems. Very little effort will be devoted to these problems because coupled cavity TWT's can be used, thus avoiding the thermal problems inherent in the helix type TWT.

Coupled-cavity TWT's must also be considered for kilowatt devices above 4 GHz. A number of coupled-cavity TWT's have been developed with 7 to 12.5 kW output power from 6 to 8 GHz for use in ground stations for satellite communication (see Table 12). These designs can be scaled in frequency, cathodes can be derated to extend lifetime by reducing power, consideration can be given to weight reduction and efficiency improvement techniques, and thermal problems can be reduced. This should result in a tube capable of deep space operation with a power output of 2 to 5 kW over frequencies from 2 to 16 GHz. In fact, there appears to be no reason (from the device standpoint) why this power output could not be extended to even higher frequencies.

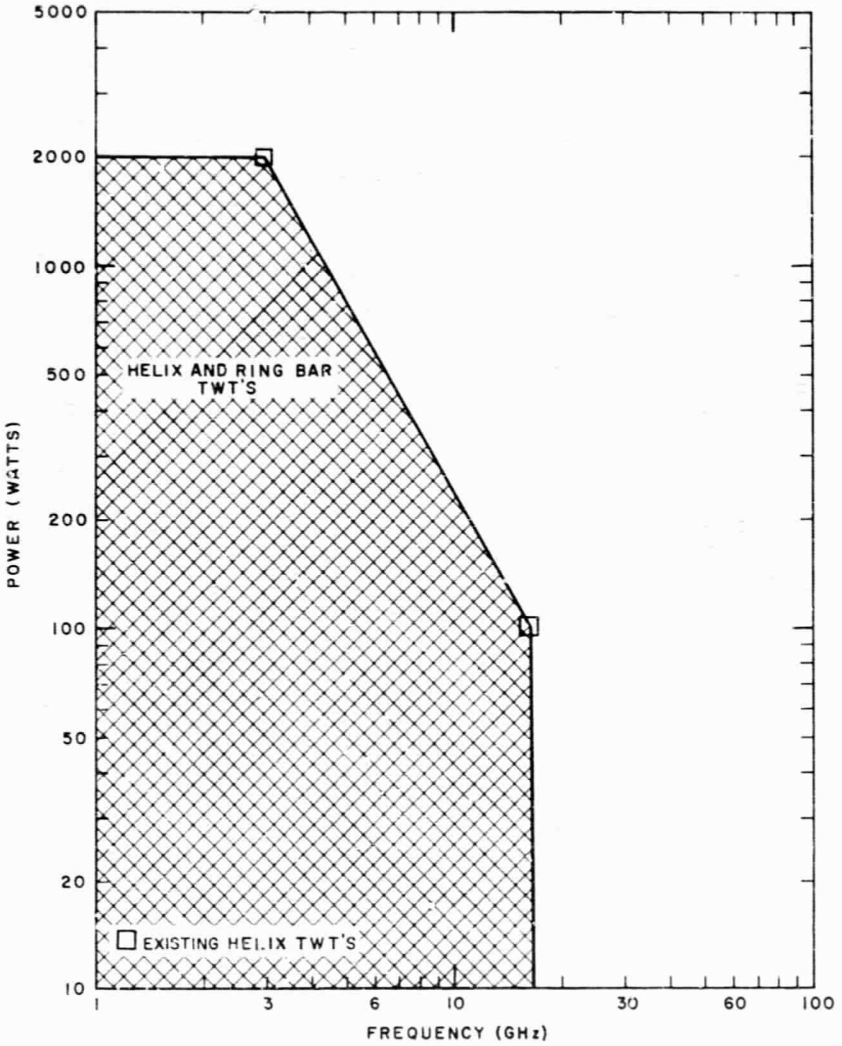


Figure 19. Limits of power and frequency for various types of traveling wave tubes

Table 12

HIGH-POWER TUBES DESIGNED FOR GROUND STATION OPERATION

Tube Type	Frequency (GHz)	Weight (lb)	Power Output (kW)
M4444	7.7–8.4	125	12.5
YH 1045	5.925–6.425	–	12.0
TWC 287	6.275–6.45	–	8.0
614 H	5.9–6.4	95	10.0
710 H	8% BW In X-band	60	10.0

The second type of linear beam device to be discussed is the klystron. This device is noted for its average power-handling capability. It is a high gain device, reliable when conservatively designed, and has well-known characteristics. Since the gun design is similar to that of the TWT, the lifetime problems should be similar.

The klystron is efficient (45 to 60 percent) at power levels from 1 to 100 kW over the microwave region of 2 to 16 GHz. (Until recently, efficiency has been poor at lower power levels because of emphasis on gain and stability.)[8] Advances in the electrostatically focused klystron (ESFK) have produced efficiencies of 30 to 45 percent.[8,16] Weight is saved by the elimination of magnets. Other advantages are that the focusing fields act as an ion trap which increases tube life, and power output levels can be varied efficiently by changing the cathode voltage.

The ESFK is a promising candidate for high power output, at low microwave frequencies in particular.[17] A 100 watt ESFK has been designed with 38 percent efficiency, 30 MHz bandwidth at 2.3 GHz, 40 dB gain, and a weight of 3 pounds. This tube is cooled by a new approach to radiation cooling – a sapphire window in the collector radiates directly into space.[13,17]

A fundamental limitation in the power capability of the ESFK is its voltage-breakdown problem. Lifetime has not yet been demonstrated, although there seems to be no reason why the lifetime should be less than that of the conventional klystron.

The highest power ESFK's known are a 1 kW tube at S band and a proposed[17] 5 kW ESFK at S band. This suggests that the kilowatt level ESFK is limited to frequencies below 5 GHz. Conventionally focused classical or extended interaction klystrons would then be required to obtain the kilowatt power level above 5 GHz.

The extended interaction klystron (EIK) makes use of a TWT type, coupled-cavity, slow wave output section. By extending the interaction between the beam and the circuit, the power per unit area of interaction surface is reduced. Thus the power-handling capability of the EIK is outstanding.[18] The EIK has yielded high efficiencies (65 percent) at the 1 kW level at S band and also for high power devices (500 kW at 8 GHz).[19–21] The bandwidth of these tubes is about 1 percent, which is approximately double that of other conventional klystrons. At present, this tube has not been developed at the lower microwave frequencies because classical klystrons have met the requirements. However, the device is currently being considered in the 20 to 500 watt range at S band for space communications.[8] There seems to be no practical limit either at lower or higher power levels. Further development remains to demonstrate long-life, reliable communications at power levels of 2 to 5 kW over the 2 to 16 GHz frequency band.

2.2 Solid-State

Solid-state sources of microwave power have often been proposed as satellite transmitters because the size, weight, and apparent reliability seem to make them obvious candidates. However, these devices lack truly high power capability with the present technology. "Present technology," however, changes rapidly in the solid-state field; the LSA oscillator diode[22] is new within the past year.

The relative capabilities and basic limitations of solid-state sources have been reviewed in a number of previous papers.[7,23,24] At low microwave frequencies (2 to 8 GHz), the multiplier chains (oscillators and transistor amplifiers at frequencies less than 500 MHz with varactor diodes to multiply the frequency) can presently obtain 10 watts or greater at relatively high efficiency (30 to 50 percent). Of the solid-state devices, only the Gunn oscillator, IMPATT, and LSA diode offer promise of yielding a higher power output.

Varactor multiplier chains may be particularly important at the lower microwave frequencies (S band). Techniques in this field are steadily advancing and the required capabilities are gradually approaching the S band frequency range. For example, a 115 W transistor amplifier with 70 percent efficiency at 250 MHz was recently reported.[25] The high power is obtained by paralleling eight 16 W power amplifiers using 3 dB hybrids. The technique is scaleable to any frequency at which a high-power transistor is available with a cutoff frequency of three or four times the operating frequency. By paralleling diodes, varactor multiplier chains are also capable of high power, high efficiency operation.[26]

A potential of better than 100 W with 30 to 40 percent efficiency exists in the near future at S band by using 500 to 1000 MHz amplifiers, then quadrupling and adding using hybrid couplers. Ultimate power and frequency limits for this type of source are not known. One of the limiting factors may be size, since typical high power hybrids at uhf frequencies are not small. An outstanding feature of this type of source would be the possibility of high reliability because of "built-in" redundancy. Such a device might be attractive as a backup to a linear beam device (TWT or klystron) to provide redundancy rather than two identical linear beam devices. For a power output equal to the primary device, the size of the multiplier chain might be excessive; however, it might be suitable as a low power backup of reasonable size.

The Gunn diode at present is limited by the lack of highly pure and well controlled GaAs material. Along with the IMPATT device, there is a low-frequency limit for CW operation in the vicinity of 3 to 10 GHz because of increasing thermal resistance of the active region. The maximum output for the Gunn device in the 2 to 10 GHz region is not expected to exceed 10 W in the near future.[27] The maximum power that can be obtained from IMPATT and Gunn devices at higher frequencies varies with frequency to the -2.5 power because the thickness of the active region, and hence the maximum voltage, must decrease as the reciprocal of the frequency. The current through these devices must decrease even faster than the voltage as the frequency is increased in order to keep the magnitude of the device negative resistance larger than the series circuit loss resistance which (because of skin effect) increases as $f^{1/2}$

The IMPATT diode now operates near its theoretical limit. The ultimate CW power of this device probably will not exceed 50 W at 10 GHz. Efficiencies of these devices are about 6 percent. It is expected that 12 percent is a reasonable estimate of the ultimate diode efficiency.[24]

The potential of the LSA diode has not been fully realized.[22] The maximum size of an LSA diode has no intrinsic limit because the LSA mode is not subject to the transit-time limitation. These larger diodes provide the possibility of bonding more efficient heat sinks to the material, thus permitting higher power to be generated. Practical limits are set, however, by the dimensions of microwave cavities and the size of uniformly doped gallium arsenide crystals.[27] To make the most of the LSA mode of oscillation at microwave frequencies, the original oscillator diode design, which is thinnest in the direction parallel to the current, must be altered so that the diodes are long in the direction parallel to the current and thin in a direction perpendicular to the current.[22] It has been predicted that power levels of the LSA devices will increase 1 to 2 orders of magnitude within a year[27] and ultimate power levels in the kilowatt range have been projected. However, some reservations have been expressed about these projections.[28] The required highly pure and well controlled GaAs material may not become available quickly. Techniques to insure adequate thermal dissipation have not been proven and reliability of these high power devices is questionable. Also, the electric field intensity, maximum current density, and thermal conductivity are lower in GaAs oscillators than they are in Si IMPATT,[28] thus the inherent CW power capabilities of LSA devices may be less than that of the IMPATT – in particular at frequencies near 10 GHz where the IMPATT is performing best. Efficiencies obtained thus far are low; the maximum efficiency predicted for GaAs is 18.5 percent.[28]

Thus, except for the multiplier chains and perhaps the LSA device, the solid-state sources have maximum power outputs which are not and evidently will not be high enough to satisfy requirements for deep space communications 10 years from today. One way to use transistors and IMPATT oscillators might be to parallel the transmitters by phase-locking them. It may be possible to do this with transistors more easily than with IMPATT diodes. Some preliminary work of this kind (with IMPATTs) has been performed,[30] although the weight and size of the resulting source would be substantially greater than that of the sum of the individual units.

Assuming that by some means (unknown at the present time) the power level of solid-state devices proves satisfactory at some future date, its low efficiency may still rule out its use. (This statement includes all solid state devices except the multiplier chains previously mentioned.) For example, if the solid-state source has an overall efficiency of 10 percent and a tube type source has an overall efficiency of 40 percent (including power supply), the weight of the prime power for the solid-state source would be higher by a factor of four and the cooling capacity required would be six times as great as that required for the tube.

2.3 Performance Characteristics

The klystron and traveling wave tubes are the most likely candidates for use as deep space transmitter tubes. It is not clear that one class is superior to the other; therefore, the performance parameters discussed pertain to both classes of tubes in general

2.3.1 Power

Klystron and TWT's are capable of at least 5 kW over the entire microwave region. A 5 kW space-qualified tube can be obtained in ten years but would require an effort substantially increased over that which now exists. In comparison, a 2 kW tube will probably not require a larger effort. The coupled cavity TWT's and the extended interaction klystron have a 2 to 5 kW power capability over the entire microwave region. Effort should be concentrated on these tubes.

2.3.2 Efficiency

High efficiencies (≈ 65 percent) have been reported using extended interaction klystrons[19-21,23] and coupled-cavity TWT's.[31,32] Problems exist in incorporating these techniques at high microwave frequencies (12 to 16 GHz); however, a goal of 50 percent over the entire microwave region is realistic. Programs presently exist to obtain efficiencies of 45 to 55 percent at S band with space qualified tubes.[12,13]

The net power efficiency of high power tubes is reduced if power is required for solenoid focusing. The solenoid for a 2 kW TWT uses approximately 600 watts. Assuming a 50 pərcent tube efficiency, the overall efficiency would be reduced to approximately 45 percent.

The choice of a permanent magnet to focus a TWT would involve an efficiency-weight trade-off since the magnet weight is higher than that of the solenoid (at least for microwave frequencies less than 8 GHz). No known efforts exist to develop lightweight permanent magnets to focus high power TWT's at low microwave frequencies. Permanent periodic magnets and field reversal magnets have been successfully used to focus microwave devices up to several hundred watts of output power. Extending the capabilities of these schemes to the kilowatt level and above will require continuing development of temperature-stable, lightweight magnets (such as platinum-cobalt) capable of high fields. Technically such magnets probably can be made. So far, however, the cost of candidate materials is prohibitive. At the present time the required capabilities do not exist and no significant effort is being made to develop them.

Extended interaction klystrons also require solenoid focusing for high-power operation over part of the microwave band. An alternative at the lower microwave frequencies is electrostatic focusing; however, ESFK power capabilities are limited. The results of the efficiency-weight trade-offs between permanent magnets and solenoids for klystrons are not known.

2.3.3 Weight

When estimating overall high power transmitter system weight, errors in predicting a relatively small weight such as the transmitter tube should be insignificant compared to the weight of the prime power source. The procedure used to estimate tube weight is described without justification. However, as a result of the procedure, a satisfactory predictor is obtained with the uncertainty insignificant compared to the total high power system weight.

Weight vs. power in each of the frequency bands from 2 to 4, 4 to 8, 8 to 12, and 12 to 16 GHz were initially plotted for CW tubes listed in manufacturer's data and in other general literature. Weights obtained include focusing solenoids or magnets but not the power supply or cooling system. Approximately 100 points in the power range from 2 to 10,000 watts were included with no attempt to bias results by any elimination process except for a few obvious cases of extremely high weight. Initially, TWT's were separated from klystrons, but no significant trends were noticed so the data were combined. The data for the different frequencies are shown on Figure 20. Also, an indication of the number of sample points from each manufacturer is given. Tubes which are either proposed or used for a space application or have characteristics similar to those required for a space environment are specially indicated. The curve shown on the figure is a minimum weight prediction for future space qualified microwave tubes, independent of frequency. Comparing the prediction with present space qualified tubes indicates that the uncertainty in the weight prediction is relatively small.

2.3.4 Lifetime

The reliable lifetime of a vacuum tube may be considerably different from the lifetime quoted by manufacturers. The quoted lifetime is usually the design life of the cathode. The reliable lifetime of a microwave vacuum-tube amplifier is actually determined by many factors, including mechanical construction, impurities in the cathode, manufacturing processes, and operating power, and can be determined only by a qualified reliability study and life test program.

Several programs have been initiated to test low-power, space-qualified traveling wave tubes and as a result a satisfactory technology for low power tubes has evolved.[33-35] Some of the information obtained in these programs can be applied in the design of high power tubes.

A coupled cavity TWT with a theoretical lifetime of 25,000 hours and a power output of 10 kW at X band has been designed and operated.[36,37] The basic electron gun design has demonstrated more than 8,000 hours of life test operation without failure.[36] By scaling this design and reducing cathode loading, a lower power tube at other microwave frequencies should have a theoretical life time greater than 25,000 hours.

Reliable life time, power output, weight, and efficiency are related in a complex manner. However, the 2 kW klystrons or TWT's previously discussed should be capable of 20,000 hours of reliable life at frequencies of 2 to 16 GHz.

2.4 Other Restrictions

There are other limitations on the transmitter power output obtainable for deep space probes. The main restriction today is the prime power source. The cost (in weight) of various types of prime power is considered in Appendix 1. Figure 21 shows the dependence of transmitter output power on weight of solar cells and distance from the sun. A transmitter efficiency of 40 percent is assumed. From a weight consideration viewpoint, a nuclear reactor would be preferred beyond 2 AU as shown in Figure 22.

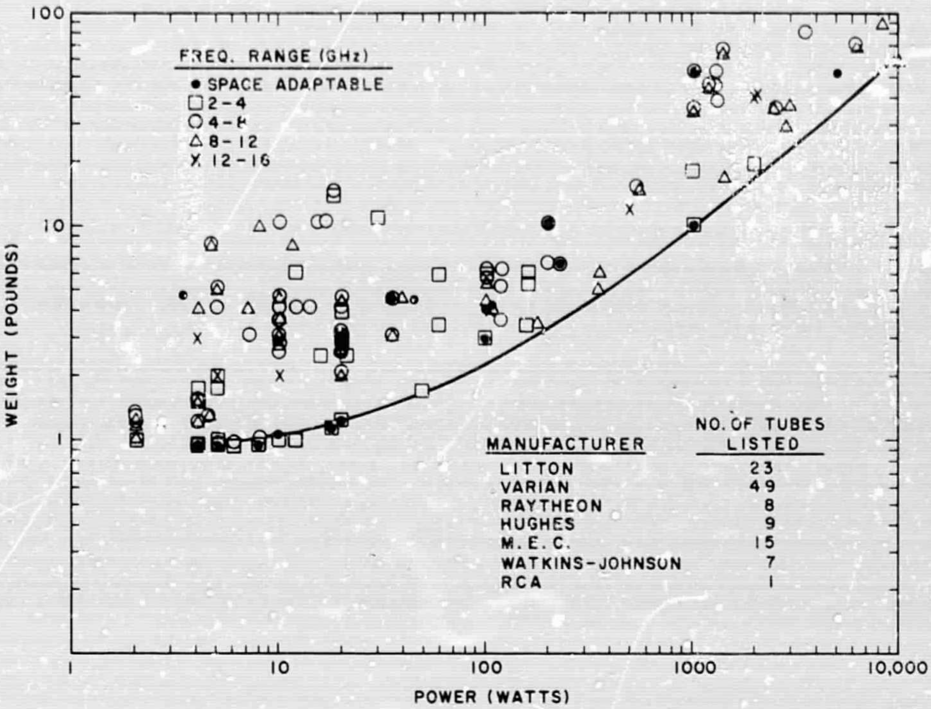


Figure 20. Weight vs. power for klystrons and TWT's at frequencies from 2 to 16 GHz

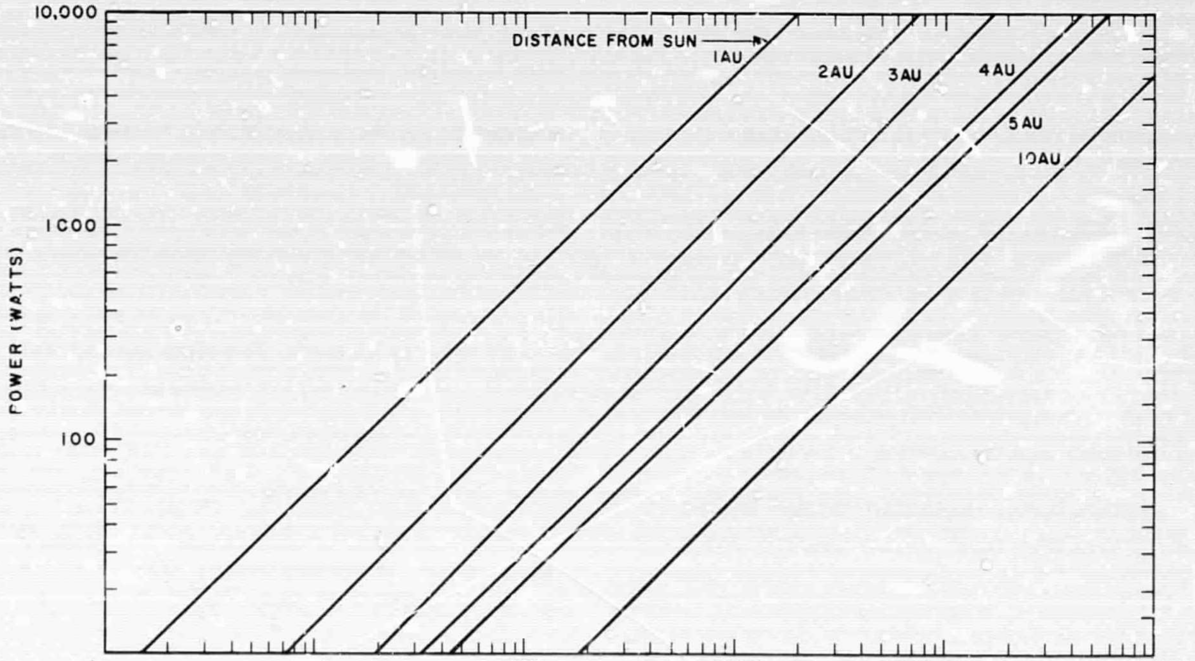


Figure 21. Weight of solar cell array vs. transmitter output with distance from sun as a parameter

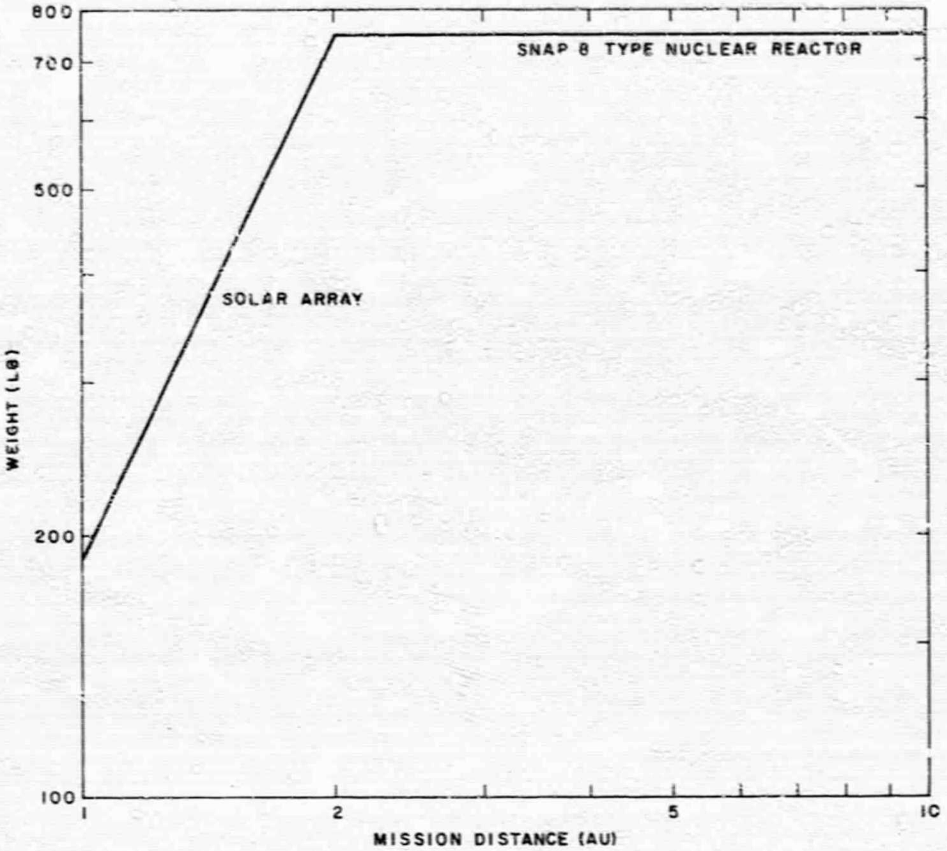


Figure 22. Minimum cost (in lb) of prime power for 1 kW transmitter output power vs. distance from sun (40 percent efficient)

The power supplies referred to in this section are associated only with the transmitter. Their function is conversion of the prime power (regulated low voltage dc) to high-voltage dc. Conversion is not required for operation of solid-state devices, but vacuum-tube devices require dc-to-dc high voltage converters.

The most important power supply characteristics for the purposes of system comparisons are efficiency, power, and weight. The best overall efficiencies reported vary from 85 to 95 percent.[38,39] Converters are less efficient at low power levels because of fixed power requirements. At higher powers the efficiencies presently obtained will not be significantly increased.

Power supplies used in space so far have been low-power devices. However, a weight-power relationship can be projected for high-power, space-qualified power supplies from data for several types of lightweight power supplies.

There have been a few high-power power supplies designed for various space applications (but not actually flown).[39-43] A few power supplies have been designed for classified, low weight, extreme environment airborne applications. Other power supplies have been built for lightweight ground applications.[44,45] Weight vs. power data for these power supplies are plotted in Figure 23.

None of the power supplies plotted precisely satisfies the requirements needed to operate a high-power transmitter. Considerable time will be required to develop the necessary power supplies. However, the data plotted provide a useful approximation of future capabilities.

An overall weight-power relationship was obtained by a straight line fit to the data. The line is biased toward power supplies that are most nearly like that required to operate a high-power transmitter. The result is given by:

$$W = 10.3\ P^{0.75} \qquad (1)$$

where W is in pounds and P is in kilowatts. In addition, the results of a 1962 internal BTL study has been added in Figure 23, thus offering a comparison with the 1962 state of art.

Integrated transmitter power supply units would weigh less than the sum of the weights of units designed separately. Such an approach has been considered.[17,46] Whenever possible it is recommended that the procedure be adapted to future designs.

Another limitation on output power is that power must be dissipated at the rate of the transmitter output power times (1 – Efficiency)/Efficiency. At present, the Mariner spacecraft has experienced some difficulty dissipating 75 watts.[47] For powers in the kilowatt region, the technology being developed to enable devices to be cooled adequately is vital.

There are several cooling techniques which offer promise: the "heat pipe" technique,[47,48,49] and the direct radiation of thermal energy into space.[18] The ultimate capabilities of these techniques are not known; however, the additional weight of these systems should be small compared to the weight of the prime power.

A limitation is also imposed on the average power rating of the output waveguide.[50] Because of the finite conductivity of the waveguide walls, power transmitted through the waveguide is attenuated. The power loss is dissipated as heat. Thus the microwave power through the waveguide is limited by the maximum allowed waveguide wall temperature.

In general, the wall temperature should not exceed the point at which the tensile strength, yield strength, and hardness begin to decrease rapidly. The temperature at which this occurs varies with the type of material but is approximately 200°C for materials commonly used for waveguide (copper, brass, or aluminum). For a standard rectangular copper waveguide in a ground environment with convection and radiation cooling, an allowable temperature rise of 110°C from an ambient 40°C limits the average power in the waveguide to approximately 1.8 kW at 16 GHz.[50] The waveguide in a typical spacecraft environment (with no active cooling) is cooled largely by radiation to the surrounding environment. Such a waveguide would be designed for minimum weight (thin walls) with maximum surface area for radiation. Conduction cooling is limited by the thinness of waveguide walls and by a large thermal impedance at structural joints when these joints are in vacuum.

As an example, consider a standard 16 GHz copper waveguide in a spacecraft environment. Assume that the ambient temperature is 30°C, the wall temperature is allowed to increase to 200°C, the cooling is by radiation only, and the radiating surface of the waveguide is increased by a factor of 10 by adding fins. By using the data presented in Reference 50, it can be shown that the power limit is approximately 4 kW for this example (at 16 GHz).

Because the parameters were chosen arbitrarily, this limit is not strict. However, the example is optimistic in that it allows the temperature to reach 200°C and does not allow a safety factor for standing waves in the waveguide[50] (which would increase the wall temperature). Therefore, this limit indicates that due consideration should be given to the exact problem. For larger waveguides at lower frequencies, the average power rating increases. It is possible that the waveguide temperature will require active thermal control at the higher power levels.

2.5 Conclusions

The most likely candidates for deep space transmitter tubes are the klystron and the traveling wave tube. Table 13 summarizes the characteristics of the future deep space tube.

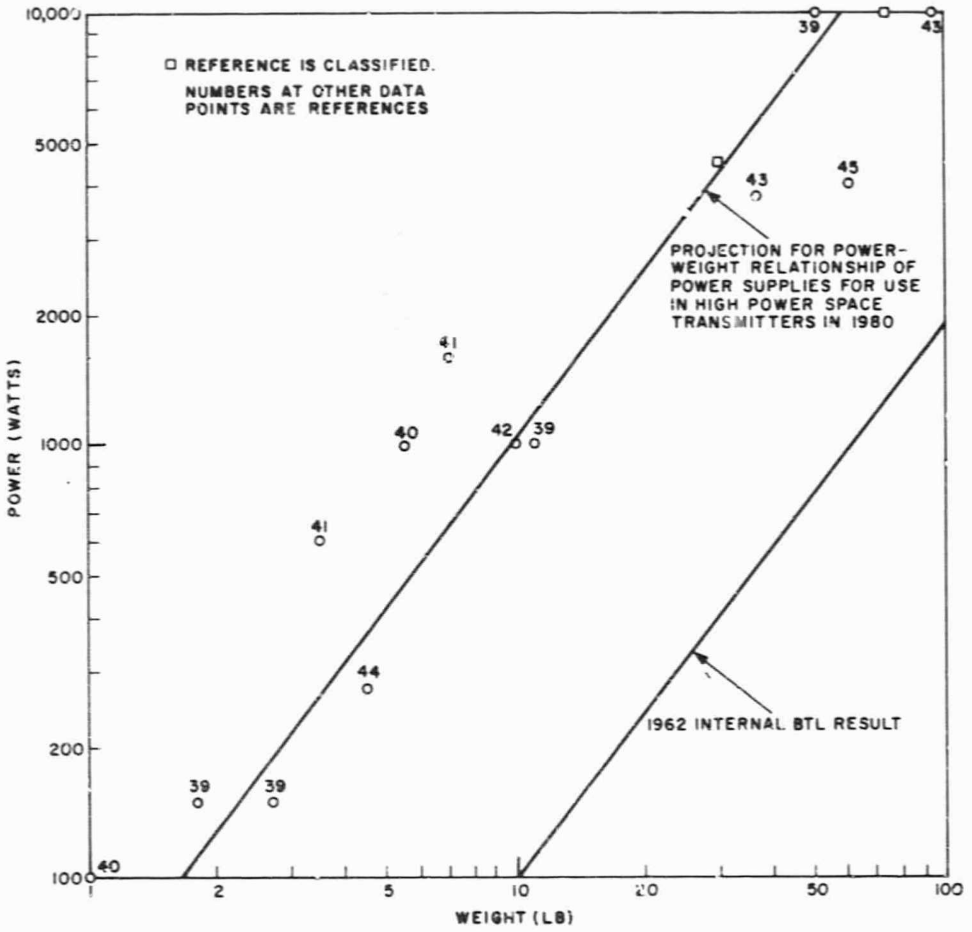


Figure 23. Weight vs. power of space adaptable and other lightweight power supplies

Table 13

CHARACTERISTICS OF TUBES FOR DEEP SPACE COMMUNICATION FROM 2 TO 16 GHZ

Lifetime	20,000 hours
Power	2 kW
Efficiency	50 percent
Weight	16 lb

Even though it is possible to attain an output power of 5 kW in a space-qualified tube, restrictions imposed by cooling requirements and the weight of the prime power tend to suggest output powers in the 2 kW range.

Means must soon be found for cooling the high-power space transmitters at microwave frequencies. The solution to the problem will result in some additional weight; however, it should be small compared to the weight of the prime power (see Appendix 1). The weight of an entire transmitter system – prime power, power supply, cooling system, and transmitter tube – can be estimated (as a function of power output, efficiency, and mission distance) using data presented in this section. Figures 24 to 26 show the results plotted for mission distances of 1.5, 2, and 2.5 AU from the Sun.

3. SPACECRAFT ANTENNAS FOR MICROWAVE AND MILLIMETER WAVE COMMUNICATIONS SYSTEMS

The field of high-gain parabolic reflectors for spacecraft is characterized by a proliferation of different structural concepts, development techniques, and promising ideas, few of which have actually been reduced to practice except in scale-model experiments. Typically, gains of 40 dB or less at gigahertz frequencies have been acceptable in operational antennas. Such antennas provide the only actual experience for this study. To investigate the potentialities for a range of frequencies from 1 to 100 GHz with antennas weighing from 50 to several hundred or even thousands of pounds, considerable reliance must be placed on proposals for these kinds of structures. A good bit of faith in the ability of industry to reduce the results of these feasibility studies to reliable or operating hardware is also necessary.

The basic performance characteristic used for comparison is power gain, just as it will be in the ground antenna study of Section 5. The available gain is considered as a function of weight and frequency. Weight replaces cost as an independent variable here, since the cost of a high-gain antenna is a relatively minor contribution to the total cost of launching a complex spacecraft. This is not to say cost can be neglected entirely, however. The total development cost of a complicated, deployable, large diameter antenna is measured in millions of dollars, and clearly a cost trade-off must eventually enter the picture. Nevertheless, for a general comparison, antenna weight is the more natural variable to use at this stage.

A diversity of structures has been proposed to meet the need for spacecraft reflector antennas, but no single concept is applicable to the range of operating frequencies, weights, tolerances, and gains that are of interest here. Five different types of antennas are considered. These types of antennas together span the entire region of interest, but there is no clear demarkation between alternative types within the region. The five types of antennas will be discussed in more detail below.

Each antenna type, or class, is characterized by two relations which express the weight and the rms surface tolerance as functions in diameter. In every case, a power-law expression of the form αD^{β} was appropriate for these relations, but the parameters α and β are different for each class. These expressions were combined with Ruze's gain formula[51] in order to relate the gain to frequency and weight. When this had been done for each class, the classes were considered together and a composite performance estimate was established by constructing a gain envelope for the performance of each antenna class considered separately.

The study concludes with a brief discussion of the problems of attitude control and antenna-pointing tolerances. It is difficult to discuss such problems in general terms, but it is clear that difficulties may arise if very high gain (60 to 70 dB) antennas are considered. In any specific case, it should be possible to generate an acceptable solution to the control problem. It is presently not possible to estimate either the weight penalty or the required development time explicitly.

Phased array spacecraft antennas are not treated in detail Past considerations regarding satellite antennas at BTL have generally led to the conclusion that the weight, cost, and complexity of phased arrays were not justified. Phased arrays are attractive when their special capabilities are needed, such as rapid and versatile beam steering, multiple beam operation, unusual aperture illuminations, or very high power levels. It seems clear that a deep space probe does not require the special features of phased arrays, and weight considerations argue strongly for conventional antennas. A study[55] by General Dynamics provides an example of the weight penalty involved. In their treatment, the phased array is more than an order of magnitude heavier than conventional types. A 60 ft diameter phased array would weigh about 25,000 pounds.

3.1 Environment, Design Loads, and Other General Considerations

The weight assumed for each class of antenna includes the weight of the primary reflector, the feed support structure, the structural attachment to the spacecraft, and

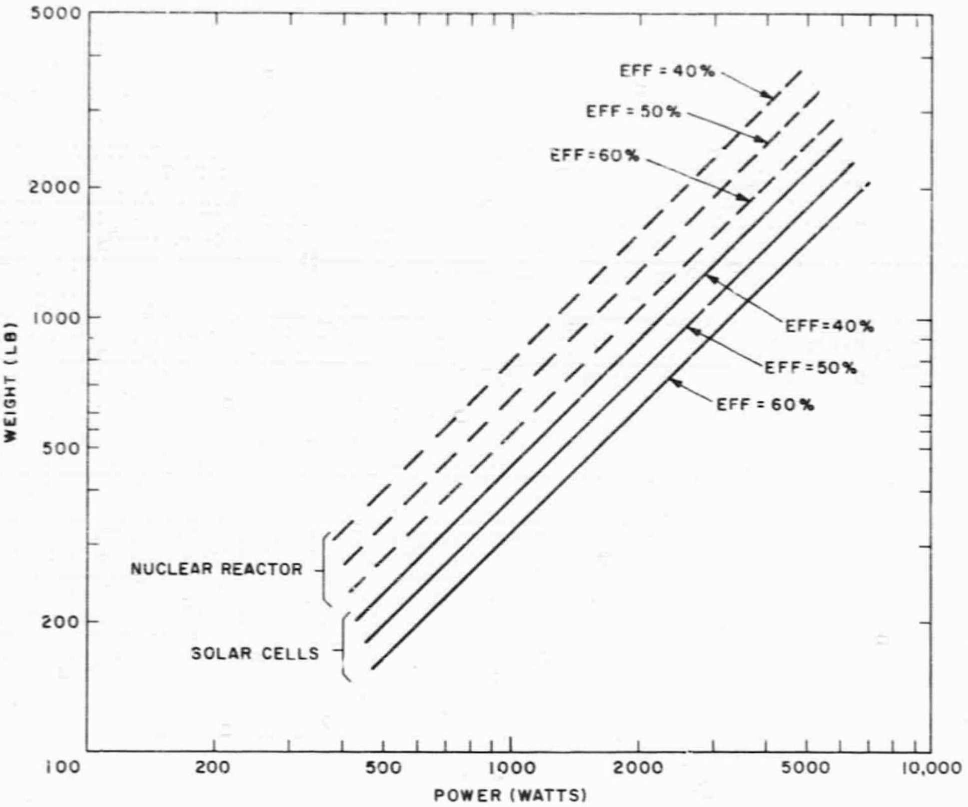


Figure 24. Weight of transmitter vs. power output with overall efficiency and type of prime power as a parameter (1.5 AU from sun)

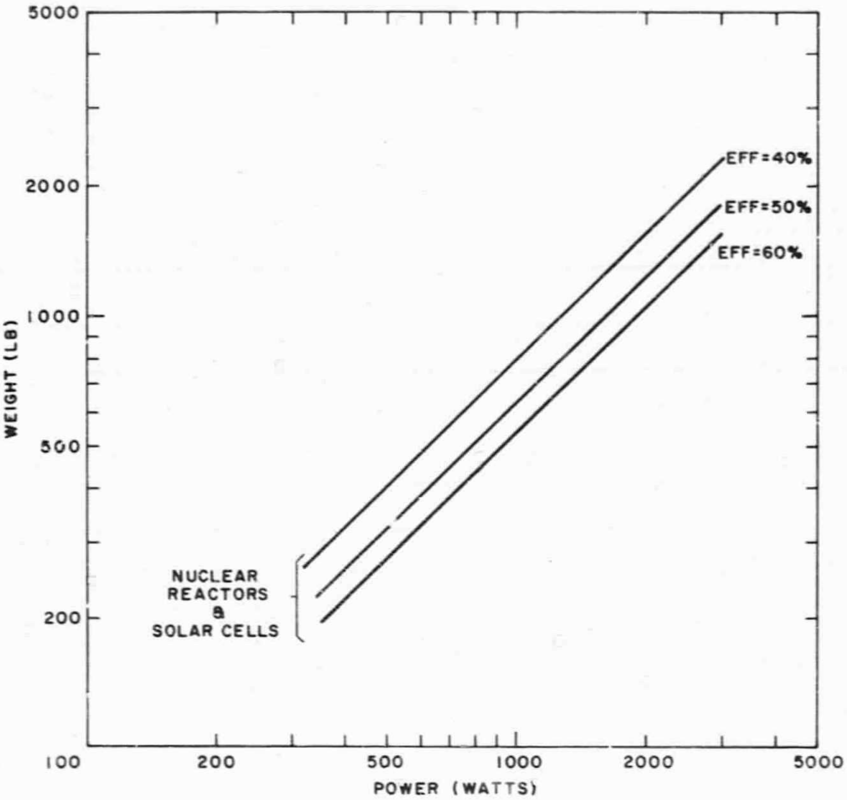


Figure 25. Weight of transmitter vs. power output with overall efficiency and type of prime power as a parameter (2 AU from sun)

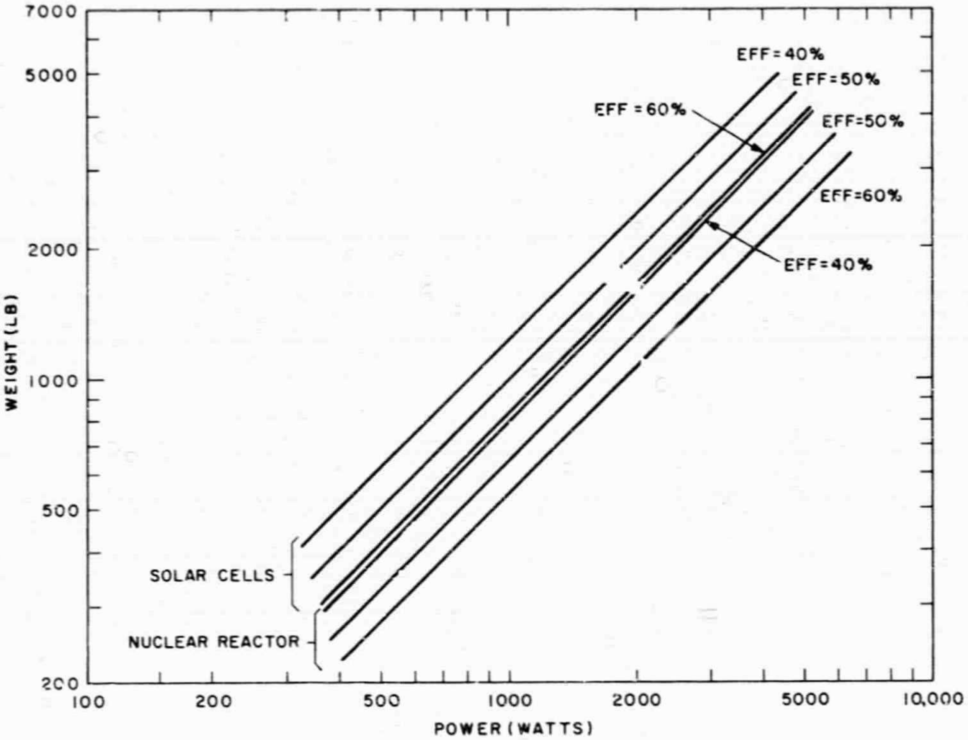


Figure 26. Weight of transmitter vs. power output with overall efficiency and type of prime power as a parameter (2.5 AU from sun)

the erection mechanism, if applicable. No provision for weight of electronics has been made. No separate consideration is given to the various types of feed arrangement and design that are possible; e.g., Cassegrainian, or focal point feed. There is insufficient information regarding the weight of such alternatives to permit any important distinction to be made. The antenna weight does depend on the ratio of focal length to diameter of the reflector, since the f/D ratio influences the surface area, and hence structure, for a reflector of diameter D. However, the f/D ratio for most spacecraft reflectors will be in the range of 0.30 to 0.40, and the variation of weight in this restricted interval is not pronounced. As a result, the dependence of weight on reflector f/D ratio is not included in this study. Cases which may require an unusually deep or shallow reflector because of special design requirements must be considered separately.

The structural loads which influence the design of the spacecraft reflector for a deep space probe are relatively benign. The thermal environment caused by solar flux and the dynamic loads associated with maneuvers of the spacecraft are the major effects to be considered. The thermal environment for a deep space probe is much different than for an Earth-orbiting satellite. The incident flux is reduced with increasing range from the Sun. In addition, as the range of the probe increases, the solar flux becomes more unidirectional with respect to an Earth-pointing reflector. For example, at 2 AU from the Sun, the Sun lies within 30 degrees of the axis of an Earth-pointing antenna. At 10 AU this angle is reduced to less than 6 degrees. This is in distinct contrast to Earth-pointing reflectors on orbiting satellites, which are exposed to solar flux from virtually all directions (relative to the antenna) during their lifetime and which also are in the Earth's shadow occasionally. This condition substantially alleviates the thermal design problem. The differential deflections which exist when a reflector is exposed first from the front and then from the back do not occur. The fact that the thermal loads act in a direction which is relatively constant with respect to the reflector suggests that their detrimental effect will be substantially reduced and perhaps eliminated entirely by proper design.

The dynamic loads are another matter. As long as maneuver of the spacecraft is necessary, which for practical purposes is throughout its entire useful lifetime, the dynamic loads will be present. Their magnitudes can be minimized by maneuvering with low accelerations over long periods of time, if such constraints are compatible with other mission requirements, but dynamic loads cannot be entirely eliminated. It is also important to realize that the deflections from dynamic loads cannot be reduced very much by providing a stiffer structure. This is especially true of shell-type reflectors, because membrane stresses in the shell are the primary means by which the loads are carried. Membrane stiffness and mass both increase linearly with thickness, so no improvement in dynamic response whatsoever is provided by a stiffer structure. If the reflector has a back-up structure of some sort in which significant bending action can be realized, some improvement in dynamic performance can be effected. The important point to bear in mind is that an effective weight-surface accuracy trade-off is much less plausible for spacecraft antennas than, for example, the cost-surface accuracy trade-off postulated for ground antennas. As a result, the rms surface accuracy of a spacecraft antenna is determined mainly by the type or class of antenna chosen rather than by design modifications within any particular antenna class.

From a digest of the existing literature, including proposals and state-of-the-art surveys, a relation between rms surface tolerance and diameter is suggested for each antenna class considered. A certain amount of faith attends such a suggestion because of the uncertainties involved and the rather skimpy data one has to go on in each case. The rms error budget for surface deviations was established by doubling the estimate for manufacturing surface tolerance achievable for each type of antenna. Even this arbitrary factor of two represents a considerable oversimplification of all the conditions involved. Deflections caused by environmental loads are generally much larger than manufacturing surface tolerances for small, high-precision reflectors; the converse is true for large, inflatable antenna structures, for example. However, the environmental loads themselves are relatively modest in the present application, as already pointed out. In any case, the factor of 2 is intended to include all deflection effects which contribute to loss of gain (such as the feed-support deflection, which is not explicitly considered otherwise). Such a factor also reduces the effect of the uncertainties inherent in the estimation of the manufacturing surface accuracy attainable for each case. This factor also recognizes the principle that if one insists on attacking areas fraught with unknown quantities, some vestige of technical respectability can be maintained by being conservative. A factor other than 2 can, of course, be incorporated without difficulty if its validity can be demonstrated.

There is often confusion about rms surface tolerances in respect to the datum to which they are referred. It is not unusual to achieve a reduction in rms surface tolerance by an order of magnitude simply by measuring it with respect to a best-fit paraboloidal surface instead of the design paraboloid. However, this procedure implies that one will be able to take full advantage of the revised datum by placing the feed at the focal point of the best-fit paraboloid. This, in turn, requires a movable feed, and it is unlikely that such sophistication will be available in spacecraft antenna systems in the near future. Hence, the tolerances suggested here are with respect to the design surface unless explicitly noted otherwise.

3.2 Gain Formula

Ruze's gain formula[51]

$$G = \eta \left(\frac{\pi D}{\lambda}\right)^2 \exp - \left(\frac{4\pi\epsilon}{\lambda}\right)^2 \qquad (2)$$

is used to describe the dependence of gain on diameter, frequency, and surface tolerance. Extensive use of this formula is also made in Section 5. D is the diameter, λ the wave length, and ε the rms surface tolerance and all are measured in the same units. The aperture efficiency is denoted by η and is set equal to 0.70 henceforth. The justification for this selection is identical to that given in Section 5.

Ruze introduces several assumptions in the derivation of Equation (2). He assumes the deviations of the surface are uniformly distributed over the aperture, and the local deviations from the desired surface are essentially independent, and obey, at least approximately, a Gaussian distribution law. Ruze gives a good discussion of the limitations of Equation (2) due to these constraints in his recent article.[51]

The significance of the rms surface tolerance in Equation (2) is apparent. It is somewhat disturbing to find such a poorly defined quantity in such a sensitive spot. It is clear why communications engineers require surface tolerances to be a small fraction of the operating wavelength. Maximum gain, for a given reflector, is achieved at that frequency for which λ = 4πε (known as the gain-limit point). Here the loss of gain due to surface imperfections is ~4.3 dB compared with a perfect reflecting surface. Beyond this point the gain is a strong function of a parameter which depends on the correlation length (see Appendix 3).

To illustrate the significance and elusiveness of the rms surface tolerance, it is worth considering two examples which lead to somewhat surprising conclusions. Mar and Wan[52] have investigated the rms surface tolerance of a shell-type reflector under gravity load with all other variables held constant. They find that ε actually increases as shell thickness increases. Although membrane deflections under this loading are independent of shell thickness, there is an annulus at the edge of the reflector in which bending effects are somewhat larger in thicker shells than in thin ones. The net effect is an increase in rms surface tolerance with thickness.

The second surprise occurs when the rms surface tolerance for a solid shell-type reflector is compared with that for an open truss-like reflector under thermal loadings from incident solar flux. Although the magnitude of the surface deviations from the design paraboloid are generally much higher for the shell than for the truss, the rms surface tolerance for the shell is often much lower if it is referred to the best-fit paraboloid. The deformations of the shell-type structure are more nearly homologous (see von Hoerner[53]) than those of the truss under these conditions. These examples illustrate the dangers inherent in attempting off-the-cuff estimates concerning the rms surface tolerance under various types of loading conditions.

The second example just discussed suggests a further deterrent to the injudicious application of Equation (2). If the surface deviations are not essentially independent, or equivalently if the correlation interval is large, the validity of Equation (2) is compromised. Unfortunately, structural deformations from environmental loadings lead to rather large correlation intervals. Such deformation patterns result in a more severe loss of gain than predicted by Ruze's formula. Since system performance is the basic criterion, the entire antenna behavior should be studied, not just that of the reflector. This study should include the effect of feed support deflections, illumination taper, correlation interval of the reflector deformation pattern due to various effects, influence of the f/D ratio, and so forth. Such an undertaking would provide a much better understanding of the relative effect of all the various factors on system performance but is much too ambitious for a general study. In lieu of such information, system performance is assumed to be described adequately by Equation (2).

3.3 Antenna Types

This section contains a brief discussion of each of five types, or classes, of antennas that could be used on a deep space mission. Most of the information is taken from two state-of-the-art surveys recently published,[54,55] and supported by contractor information furnished with the ATS-4 proposals from several manufacturers.[56-58] For all cases, the rms surface tolerance ε is quoted in millimeters, while diameter is expressed in feet. This is for convenience in the use of Equation (2), where λ is usually expressed in millimeters.

3.3.1 Type 1: One Piece Solid Surface

These antennas are generally comparable to ground antennas at the lower end of the diameter range. A solid-surface, one-piece reflector is used. Such a reflector can be machined to a very high accuracy and offers considerable integral rigidity for the space environment even without extensive back-up structure. The maximum diameter of this kind of reflector is limited to about 25 feet, both by manufacturing constraints and launch vehicle volume budgets. It should be possible to achieve rms surface tolerances for this type of reflector equivalent to that predicted for ground antennas in Section 5. When this value is doubled, as discussed above, to provide a total error budget, the appropriate relation is $\epsilon = 2(10)^{-3} D^{3/2}$ (ε in mm, D in feet).

The weight estimate is generated by a straightforward calculation of the amount of material required by a monocoque structure of this type. The surface area of a paraboloid with f/D of 0.30 – 0.40 is approximately D^2. The ratio of the radius of curvature to thickness for such a shell is probably no larger than 500. The volume calculated using these estimates is increased by 15 percent to provide for an edge stiffener ring, feed support structure, and other miscellaneous hardware. The total weight depends on the density of the material chosen. The deformations depend on the ratio E/ρ where E is Young's modulus and ρ is the density. They are roughly independent of material because this ratio is nearly the same for all common materials. An exception is beryllium, but weight saving would probably be offset by machining difficulty in this case. If thermal distortions are a problem, a trade-off exists between materials which are relatively heavy but have a low coefficient of thermal expansion (e.g., invar) and materials which are lighter but are more sensitive to thermal loading (such as aluminum). To make a sensible decision in such cases, a detailed study would be necessary. The weight relation used in this study for this kind of antenna is $W \cong 0.3D^3$, which is appropriate for aluminum. In this relation, W is in pounds and D is in feet, a convention which will be followed throughout this section.

3.3.2 Type 2: One Piece Rigid

Antennas of this type are similar to type 1 antennas, except that they are considerably less substantial. The reflecting surface is generally mesh or perforated aluminum honeycomb sandwich material. Fiberglass reflectors with conducting material either embedded in or deposited on the surface are included in this class. A skeletal back-up structure of radial ribs is usually employed to support the reflecting surface. The diameter of this type of antenna is limited to 25 to 30 feet, mainly because of launch vehicle constraints. Although the rms surface tolerance of this kind of antenna is poorer than that of the type 1 antenna, nevertheless they can be fabricated, assembled, and adjusted prior to launch. Surprisingly good numbers are cited in the literature, in spite of the rather flimsy appearance of the structure. The relation suggested here is $\epsilon \cong 4(10)^{-3}\ D^{3/2}$, ($\epsilon$ in mm, D in feet).

A further surprise emerges in the weight estimation for this kind of antenna. A few reliable points exist since most actual experience to date has been with antennas of this class, e.g., Mariner, Voyager. The relation which seems to fit the data is $W \cong 10D$, and this includes a substantial allowance for ancillary structure such as feed support, attachment to the spacecraft, and launch reinforcement. The extremely low volumetric density of these structures is apparent. Weight scaling with diameter suggests that most of the weight is concentrated in the back-up structure rather than the reflecting surface. It is also indicative of an evolving technology which continues to devise means of squeezing the last ounce of performance out of a restricted weight budget.

3.3.3 Type 3: Petaline

Petaline is a generic term used to describe deployable antennas that resemble the petals of a flower as they deploy. There are many variants on this general scheme, some of which bear little similarity to any known horticultural species (Lockheed's flexrib deploys by rotation about the reflector axis), but they are nevertheless included here. Figure 27 illustrates a typical petaline concept.

The individual elements, or petals, of this kind of structure can probably be shaped as well as the reflector surface of a type 2 antenna. However, additional errors are introduced in the latch-up process that interconnects all the petals as the final step in deployment. Final adjustment of the deployed surface is not possible. All proposals for the 30 ft diameter ATS-4 antenna[56 – 58] suggested petaline concepts, and although the specific proposals were significantly different, there was a remarkable uniformity for both rms surface tolerance and weight estimates throughout the proposals. On the basis of this evidence, the relations used here are $\epsilon \cong 0.12D$ (ϵ in mm, D in feet) and $W \cong 0.35D^2$. The weight estimate includes the mechanism required for deployment.

Antennas of this type are appropriate in a diameter range of 20 to 60 feet. They could no doubt be made smaller, but in such sizes they would suffer in comparison with type 2 antennas. The upper limit on antenna diameter is established by launch vehicle dimensions, since the length of vehicle required (for most schemes) is approximately equal to the reflector radius. Working scale models of petaline antennas have been built and tested, and the results have established the feasibility of the concept.

3.3.4 Type 4: Expandable Truss

The expandable truss antenna has been proposed by the Convair Division of General Dynamics[55] as suitable for spacecraft applications up to 150 feet in diameter. In this concept the antenna operates like a two-dimensional compound scissors and forms a substantial back-up truss in the deployed condition. The individual elements that form the truss are usually hinged at the joints and in the center. These joints are spring-loaded to provide deployment forces. The reflecting surface is attached to the truss at numerous tie-down points, and a guy wire system is used to stretch the surface to a nearly paraboloidal contour in the deployed state. The reflecting surface proposed is flexible and folds with the truss to the stowed position. Figures 28 and 29 illustrate the expandable truss concept, and a complete description can be found in Reference 55. Working scale models of the device have been built and tested, with encouraging results.

The only source of detailed information on reflectors of this type is contained in the Convair proposal. Their

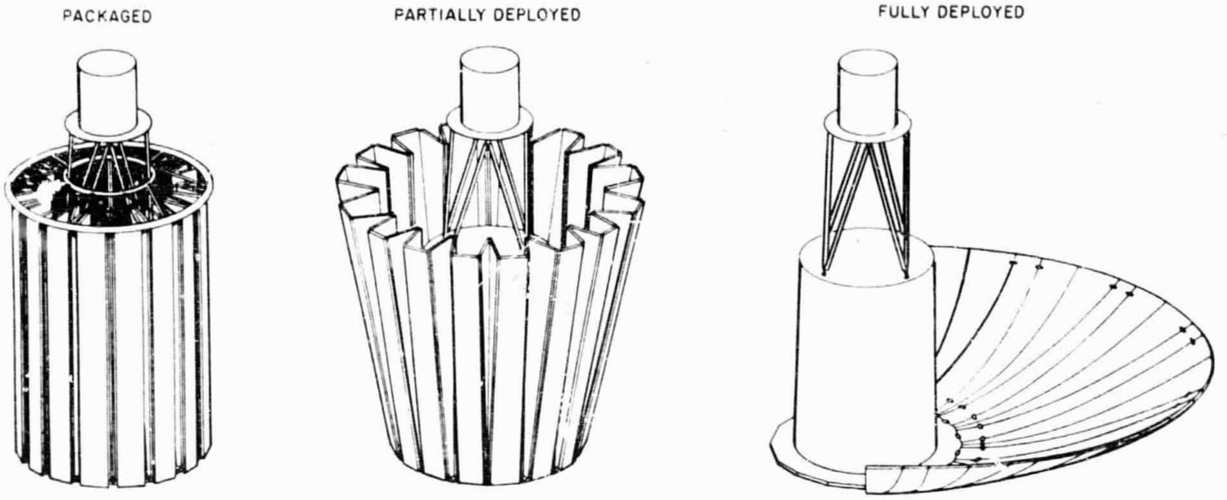


Figure 27. Typical petaline concept

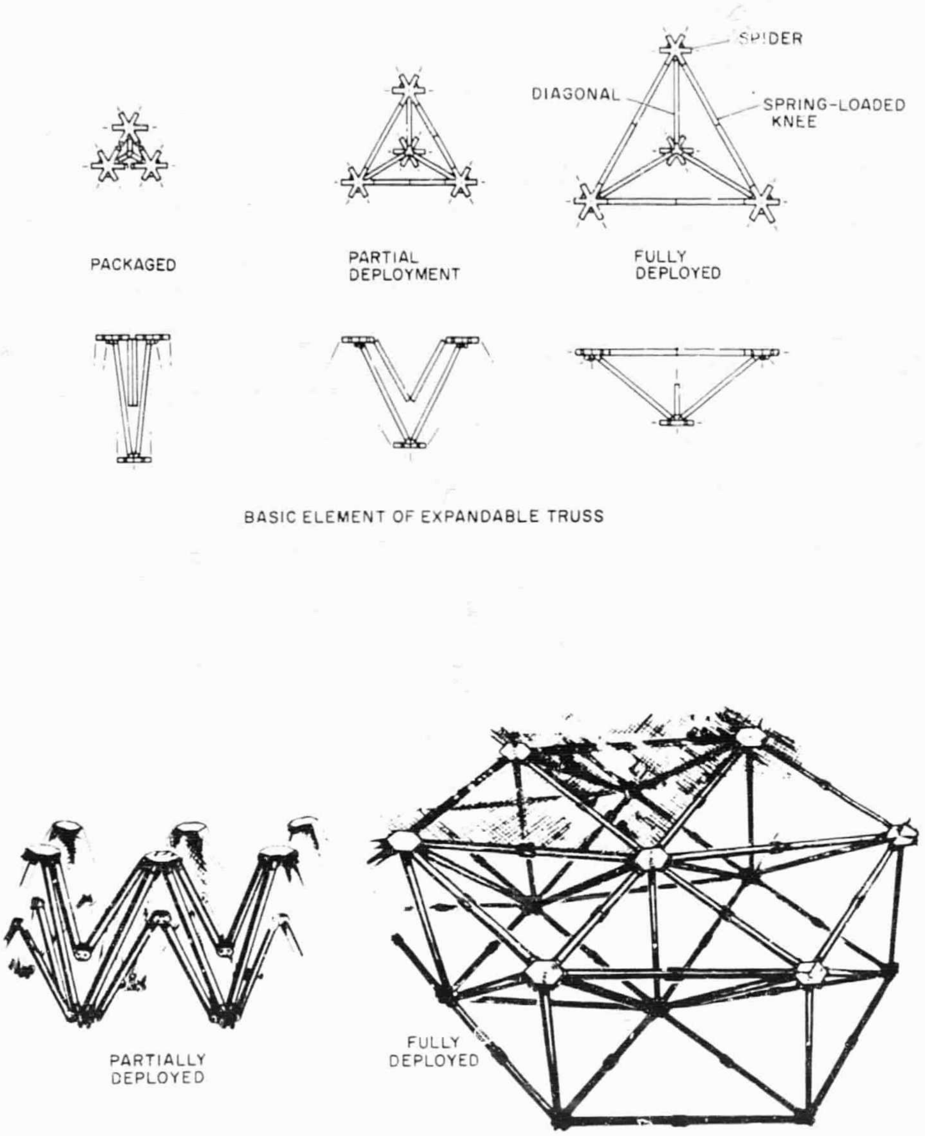


Figure 28. Expandable truss concept

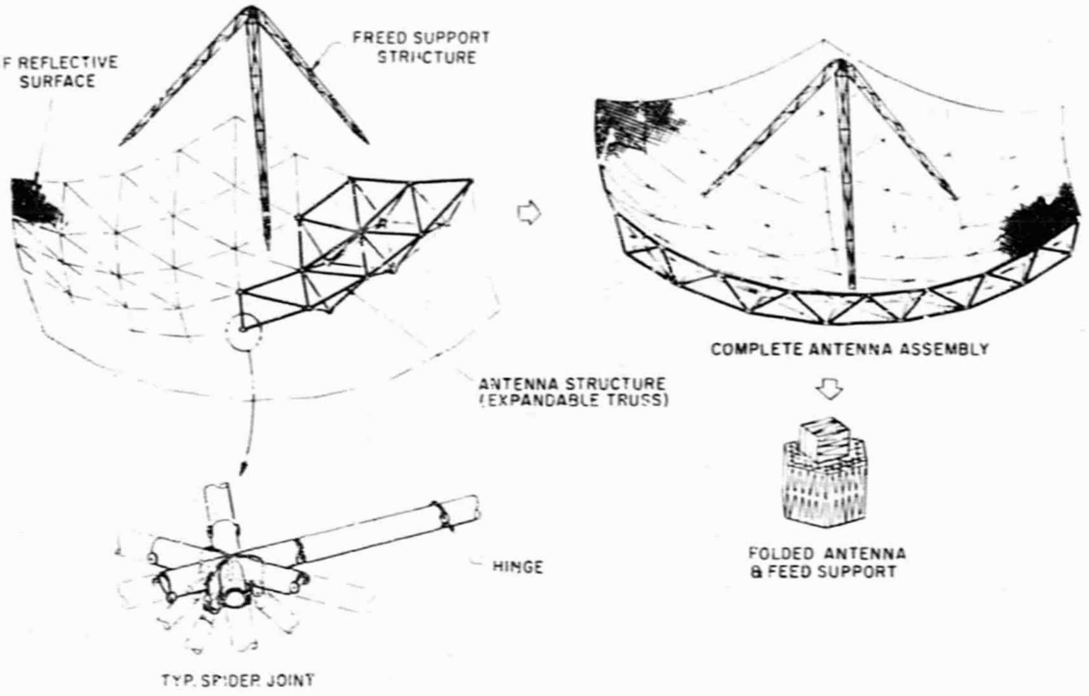

RF REFLECTIVE SURFACE
FREED SUPPORT STRUCTURE
ANTENNA STRUCTURE (EXPANDABLE TRUSS)
COMPLETE ANTENNA ASSEMBLY
FOLDED ANTENNA & FEED SUPPORT
HINGE
TYP. SPIDER JOINT

Figure 29. Expandable truss antenna

study seems thorough and conscientious, hence their suggestions for the weight and surface tolerance performance of this kind of structure are accepted here. The relations are $\epsilon \cong 0.06D$ (ϵ in mm, D in feet) and $W \cong 0.19D^2$. As mentioned above, the concept has been suggested for diameters as large as 150 feet, and there is no clear limitation that would prohibit even larger diameters. At the other end of the scale, a working model 7 feet in diameter has been built, and this is a perfectly good antenna in its own right. However, the only reason to prefer deployable antennas over type 1 or type 2 in this size range would be a severe restriction on available package volume.

3.3.5 Type 5: Inflatable

This type of antenna is included mainly for purposes of comparison and completeness. The basic antenna geometry is established by internal pressurization of certain components, and this shape can be stabilized by a variety of techniques: rigidizing foams and coatings, strain hardening of metal foils, and continued pressurization, to name a few. Because it is necessary to inflate the basic structure, the reflecting surface generally forms only a minor part of the total structure required. Some concepts suggest getting rid of this excess structure once deployment has been completed, but the total weight still must be carried aloft. Hence such antennas often turn out to be surprisingly heavy. The weight relation used here is $W \cong D^{3/2}$, which implies a weight penalty for small diameter antennas of this kind since the accessory weight they must carry for erection and stabilization is not a strong function of diameter. Only for large diameters do the weights associated with this class of reflectors look attractive.

The surface tolerance of an antenna of this class is quite poor. The reflecting surface is developed by the deployment of other members. The development of an accurate surface depends upon extremely precise manufacturing and deployment techniques which to date simply have not been realized. Small errors in inflation pressure can lead to large surface deviations. Improper curing of a rigidizing material can distort the reflector. Very little control can be exercised over these critical processes in space. The appropriate rms surface tolerance estimate for this class of antennas is $\epsilon \cong 0.6D$ (ϵ in mm, D in feet) ,and this represents an average over several different approaches which fall within this class.

Table 14

WEIGHT-DIAMETER RELATION $\left(W_i = a_i D^{n_i}\right)$ AND RMS SURFACE TOLERANCE-DIAMETER RELATION $\left(\epsilon_i = \beta_i D^{m_i}\right)$

	a_i	n_i	β_i	m_i
Type 1	0.3	3	$2(10)^{-3}$	3/2
Type 2	10	1	$4(10)^{-3}$	3/2
Type 3	0.35	2	0.12	1
Type 4	0.19	2	0.06	1
Type 5	1	3/2	0.6	1

3.4 Composite Analysis

The weight and rms surface tolerance vs. diameter relations for each type of antenna can be written in the form

$$W_i = a_i D^{n_i} \tag{3}$$

and

$$\epsilon_i = \beta_i D^{m_i} \tag{4}$$

Table 14 summarizes the parameters chosen to represent the weight and rms surface tolerance relationships for each of the five classes of antennas discussed. The parameters a_i, β_i, n_i and m_i are listed for each class, i. Equation (2) can be written

$$G = k_1 D^2 \Omega^2 \exp\text{-}(k_2^2 \Omega^2 \epsilon^2) \tag{5}$$

Here, k_1 and k_2 are constants, and Ω is the frequency in gigahertz. The rms surface tolerance is given in terms of diameter by Equation (4) and, in turn, the diameter can be expressed in terms of weight by using Equation (3). The result of this manipulation is

$$G_i = k_1 \left(\frac{W_i}{a_i}\right)^{2/n_i} \Omega^2 \exp - \left[k_2^2 \beta_i^2 \Omega^2 \left(\frac{W_i}{a_i}\right)^{2m_i/n_i}\right] \tag{6}$$

Equation (6) expresses the gain available as a function of two independent variables, frequency and weight, for each class of reflector. For a fixed weight, there is a frequency at which the gain is a maximum. This is the well-known gain-limit point. (The analysis is limited to frequencies up to the gain-limit point; see Appendix 3.) Similarly, for a fixed frequency, there is also a weight at which the gain is a maximum. This weight is given by

$$W_i = \frac{a_i}{(k_2^2 \Omega^2 \beta_i^2 m_i)^{n_i/2m_i}} \tag{7}$$

If m_i is unity, it is interesting to note that the gain at this point is independent of frequency. Also, for $m_i = 1$, the gain formula exhibits an absolute maximum in that the two maximizing conditions obtained by equating both the partial derivatives of Equation (6) to zero are satisfied simultaneously. This is not true for any other value of m_i; hence, in such cases, the frequency gain-limit and weight gain-limit points identified above are only relative maxima. An expression for the gain per pound can be obtained by dividing Equation (6) by W_i. When this relation is examined, it is found that the same maxima occur as before, except that the weight gain-limit point exists only for $n_i > 2$. Otherwise, gain per pound varies monotonically with weight.

Such points as these are primarily of academic interest here. Of more importance is the kind of performance that can be obtained, using existing and projected state-of-the-art devices, within the range of weights and frequencies specified. For each reflector class, Equation (6) defines a surface in G, W, Ω space. The surface properties are determined by the parameters listed in Table 14. If this surface is generated for each of the five types of antennas, and then the five surfaces superimposed with common axes, a composite gain surface will emerge. The surface contains several discontinuities, as constructed, but a smooth envelope can be generated without difficulty, assuming that technological developments will tend to fill in the valleys. This surface represents the best performance available for any combination of weight and frequency.

The performance surface can be represented by gain contour lines in W, Ω space. Figure 30 is such a representation. These contour lines are the result of smoothing the raw gain surface generated by the superposition. The area of the figure in which a certain antenna type is represented is not sharply defined. For weight less than ~300 lb, type 2 antennas are superior. For weights greater than ~300 lb, but gains less than ~55 dB, type 4 antennas should be used. The remainder of the figure, W > 300 lb, G > 55 dB, represents the performance of type 1 reflectors. Figure 31 gives the diameter vs. weight for each of the three types contributing to Figure 30. Hence, by using the two figures together, it is possible to determine the parameters of the antenna required for any specified level of performance, or vice versa.

Although it would be possible to derive analytic expressions for each of the contours in Figure 30, and possibly for the entire surface, this is deemed to be neither necessary nor advisable. Considerable interpolation is necessary to use Figure 30, and this is deliberate. The interpolation errors which will be made reading Figure 30 are consistent with the smoothing errors made in generating it. No further increase in accuracy, as would be implied by a complicated analytical formula, is warranted and, for system trade-off comparisons, the figure is considerably easier to use than an analytic expression.

Smith[59] suggests a relation for the weight of a spacecraft antenna which depends on diameter and frequency. His results can be converted to the parameters of Figure 30 by using Figures 2 and 4 of his paper. This comparison is shown in Figure 32. The agreement is remarkably good in view of the different approach in the two studies and the fundamental uncertainties involved. Smith restricted his study to antennas weighing 400 lb or less, but his results have been extrapolated for the purposes of comparison. Since neither his results nor the results of the present study can be interpreted with an accuracy of more than ±3 dB with any confidence, the qualitative agreement of the studies is good indeed.

3.5 Attitude Control and Antenna Pointing Requirements

The beamwidth between half-power points for a high-gain paraboloidal antenna can be estimated using the formula

$$\theta^2 \cong \frac{2.7 \times 10^4}{G} \qquad (8)$$

Here θ is the beamwidth in degrees and G is the gain in absolute units. For a 70 dB antenna, the beamwidth is approximately 1 milliradian. At a range of 1 AU, such a beam would span a diameter roughly twice the diameter of an earth synchronous orbit. The diameter spanned by the main beam of the spacecraft antenna would increase as range increased or gain decreased. Thus, at ranges of 1 AU and greater, the entire orbit of a synchronous satellite would be illuminated by the main lobe of even a high-gain spacecraft antenna.

It would be reasonable to require pointing control tolerances of 1/10 of the beamwidth. For the most demanding case, this implies a pointing accuracy of ~ 0.1 milliradian or about 20 arc seconds. The high-gain antennas discussed here will almost certainly require drive and control in two axes relative to the spacecraft, if only to isolate the other experiments and sensors from the pointing constraints placed on the communications system. Devices capable of providing position control in two axes well beyond these tolerances are available for industrial applications. Such devices would have to be adapted for reliable operation in the space environment but, in any case, meeting a requirement for relative pointing tolerances of 20 arc seconds or greater should not present severe technological problems.

The pointing tolerance constraint also introduces a requirement for attitude stabilization of the probe in an inertial reference frame. Certainly, the probe will have to be equipped with some kind of attitude control system regardless of the communications constraint. Whether this system is adequate for the pointing requirement or not requires a consideration of specific cases. For attitude control to within about 0.5 degree or better, a gas jet system may not be adequate, and gyro control will be necessary. In principle, a gyro control system can provide accuracies on the order of 1 arc second. A weight penalty is exacted in order to achieve such accuracies, but it is not a dramatic one. Larger and/or faster gyros would be required, implying weight increases in the gyro itself or the power supply system. However, a weight increase of more than a factor of 3 to 5 over a basic gyro control system would be

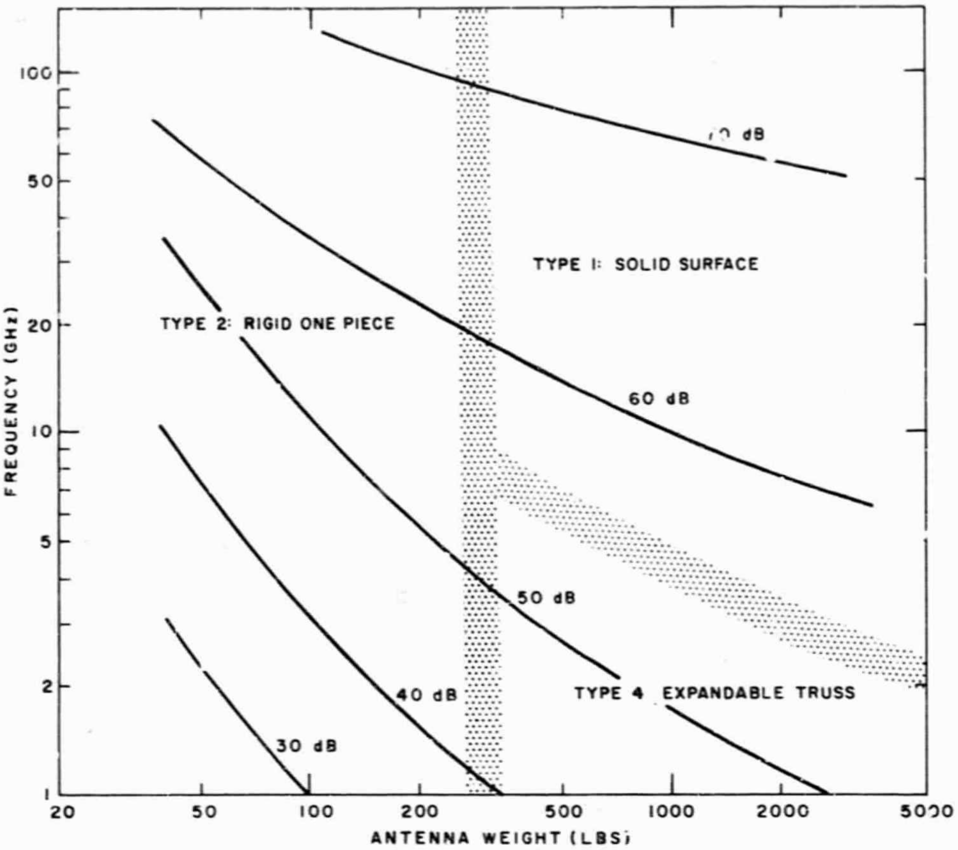


Figure 30. Composite-spacecraft antenna gain vs. weight and frequency

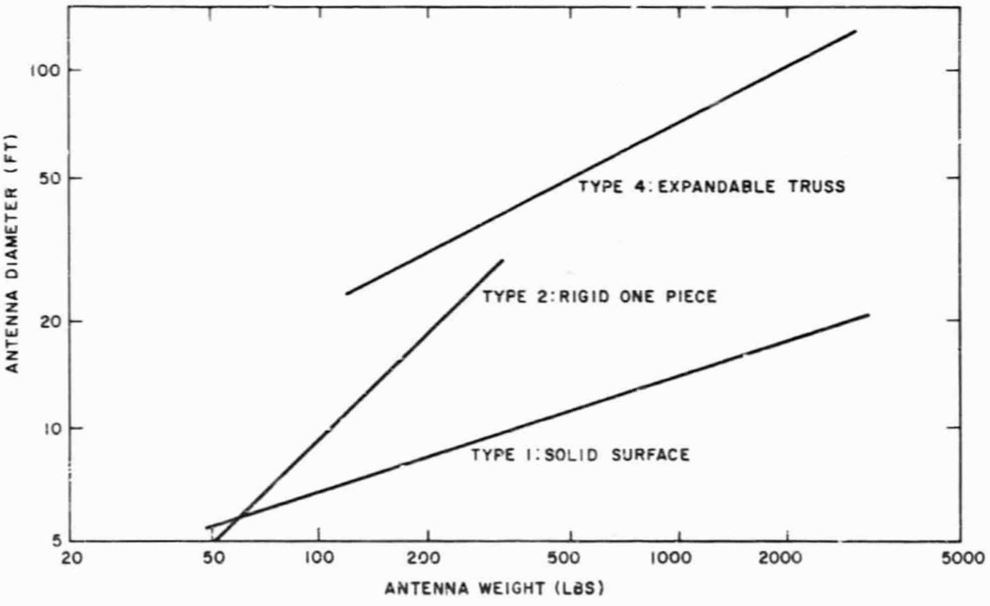


Figure 31. Diameter vs. weight for antenna types in figure 30

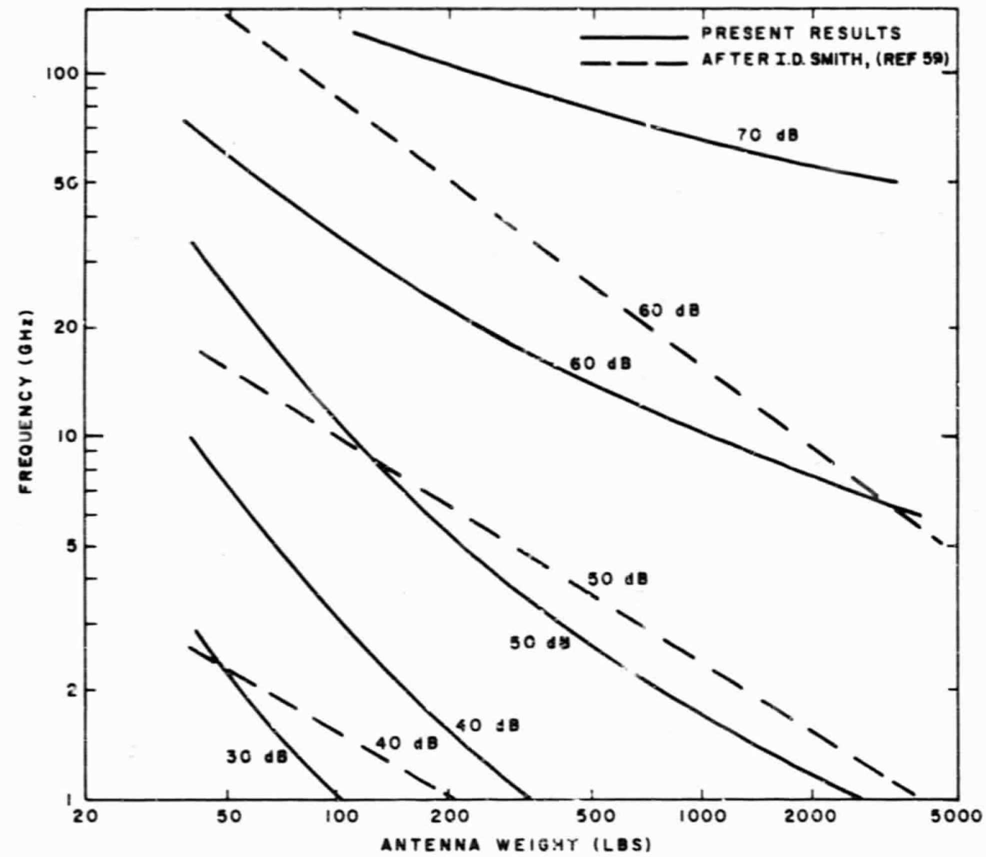


Figure 32. Comparison of spacecraft antenna gain

surprising, and since the control system is generally a very small fraction of total spacecraft weight, this increase would not be drastic. Again, particular cases would have to be considered in order to identify the specific penalties and trade-off possibilities involved. In addition, the antenna-pointing system and attitude control system are dynamically coupled. This is particularly true in the case of a heavy, large-diameter antenna. This situation would require careful study in order to establish the pointing tolerance that could be expected and the weight penalty necessary to achieve it. Other than to emphasize the necessity for careful study of the antenna drive-attitude control interface, there is little more that can be said in general terms about this problem area. The anticipated requirements, based on present spacecraft antenna performance estimates, suggest that problems may occur, but that they should not be so severe as to compromise the mission objectives if careful and thorough design practices are followed.

3.6 Conclusion

An estimate of available performance of spacecraft antennas in the frequency range of 1 to 100 GHz has been made by considering five different kinds of antennas and establishing a composite performance surface. The validity of this estimate depends on the applicability of Ruze's gain formula (which has some obvious shortcomings) and on the reliability of the weight and surface tolerance parameters associated with each antenna class. In view of these uncertainties, the present estimate is certainly no better than ±3 dB and may be somewhat worse.

Type 5 antennas, the inflatable class, need not be considered for the frequency range of interest. Their performance is surpassed by other types of antennas everywhere within the region. Type 1 antennas, the one-piece solid-surface category, are almost too good to be considered. It turns out that such antennas are necessary only if considerable gain ($G > 55$ dB) is required and one is forced by frequency constraints to expend a substantial weight to achieve it. Otherwise, a type 2 antenna will provide the required performance. Type 1 antennas are not gain-limited anywhere in the region of interest. Type 2 antennas are gain limited at approximately 67 dB, and type 4 antennas at roughly 57 dB. The occurrence of a gain-limit for these two classes of antennas dictates the transition to class 1 antennas in the upper right-hand portion of Figure 30.

It is clear from Table 14 that the expandable truss antennas (of type 4) are more accurate and lighter than petaline antennas at all diameters. As a consequence, none of the composite gain surface of Figure 30 is associated with petaline concepts. However, due to the uncertainties in the estimation of the appropriate parameters, it is entirely possible that a specific petaline design could be more accurate than, and even lighter than, a corresponding expandable truss design. Since neither class has been verified in practice in the diameter range of interest, it is probably more realistic to lump them together in a single category of large, deployable antennas, and to use the parameters suggested for type 4 antennas to represent this entire category.

Certainly, careful consideration will have to be given to the attitude control problem. Stabilizing a large diameter, high-gain reflector weighing hundreds of pounds with a tolerance of a few tens of seconds of arc will present a serious design problem but one that should be surmountable.

An interesting hybrid antenna concept can be generated by imagining a melding of type 1 and type 3 antennas. The petaline approach could be used to extend the diameter of a solid-surface one-piece central hub. This would provide a reflector with a high-precision central portion surrounded by a region with larger surface deviations, which would improve performance over that achieved with the central portion alone. Such a reflector might be particularly attractive for broadband operation. The evaluation of such a device, comprising two essentially different regions, would require modifying the gain formula, relaxing the assumptions of small correlation interval and uniformly distributed errors, and including the illumination taper. These refinements, omitted in the present study, should be included in any case in order to evaluate properly the entire antenna's performance and not just that of the reflector surface. Pending such revisions, the present approach represents a yardstick for comparative studies, in spite of the assumptions involved.

4. ATMOSPHERIC PROPAGATION EFFECTS

4.1 Attenuation and Noise

The advantages in going to higher microwave and millimeter frequencies are that larger bandwidths are available, there is less interference from overcrowding of the frequency space, and, especially significant for present purposes, it is possible to obtain larger antenna gain. An important disadvantage is that degradation by atmospheric effects increases steadily as frequency is increased. The frequency range from 2 to 8 GHz has been used widely both for Earth-based microwave radio relay systems and for satellite communications. Indeed, this centimeter band is now exhausted in some heavily populated areas.

The reason for this popularity is that atmospheric propagation loss and noise are small in this band but increase rapidly beyond 8 GHz. Atmospheric gases and liquid water exhibit strong absorption at frequencies above 8 GHz. This absorption produces attenuation of the transmitted signal and noise by spontaneous emission. The characteristics of the gases are reasonably well understood. Oxygen has a strong absorption centered at 60 GHz. Since the concentration of oxygen is stable with time, the magnitude of oxygen absorption and noise can be predicted reliably. Water vapor absorption has a maximum at 22.5 GHz which, however, is not as strong as the oxygen absorption. The attenuation and noise resulting from water vapor depend on absolute humidity, so that there are large seasonal and geographical variations. The reduction in the partial pressure of both gases with increasing altitude, of course, results in lower attenuation and noise at higher altitudes.

The lack of detailed knowledge of rain and cloud characteristics constitutes the real difficulty in predicting total attenuation and noise. Little is known about the distribution of clouds. The horizontal distribution of rain is not well known, although measurements of the kind needed are now being made,[60 - 62] and knowledge is improving. However, what matters in space communications is the vertical distribution of rain, and here almost nothing is known. Moreover, attenuation from water drops is caused by both scattering and absorption, and the magnitude depends critically on the ratio of drop size to wavelength. Attenuation thus depends on drop size distribution. This is reasonably well known at the surface of the earth,[63 - 65] but it is not known at higher altitudes, and there is no reason to expect it to be the same there as at the surface. For clouds, drop size distributions are not a question since the drops are generally much smaller than microwave wavelengths, and attenuation is simply proportional to liquid water content. Realistic calculations therefore can be made.

4.1.1 Clear-Weather Conditions

The causes of attenuation and noise now will be examined in detail, beginning with clear-weather conditions. A model which represents the typical condition of the clear atmosphere at mid-latitude is shown in Figure 33. Using this model, Hogg has computed[66] the combined effects of oxygen and water vapor. The calculated one-way attenuation through the total atmosphere is shown in Figure 34, along with measured values of absorption, and there is good agreement. (Most of the available measured data appear in a forthcoming review article.[67]) Sky temperatures are shown in Figure 35, and again the data agree well with the calculation. Both absorption and sky temperature increase rapidly beyond about 8 GHz owing to the water vapor peak at 22.5 GHz.

Large variations in the appearance of the attenuation and noise spectra are to be expected with variations in absolute humidity. The situation for very dry air is shown in Figures 36 (attenuation) and 37 (sky temperature). Here water vapor is neglected. There is little variation with frequency up to about 16 GHz. Seasonal variations in humidity can be expected to produce a 20 to 1 variation in density of water vapor, and even larger changes result when geographical differences are included.

The figures present families of curves with zenith angle (the angle between receiver axis and the zenith) as a parameter. Attenuation and sky temperature vary approximately with the secant of the zenith angle down to zenith angles of about 85 degrees.

4.1.2 Clouds

Since attenuation by clouds is simply proportional to their liquid water content, absorption coefficients and sky temperatures can be computed from the temperature-dependent and frequency-dependent complex refractive index.[68] Attenuation is shown in Figure 38 for temperatures of 0°C and 20°C. Use of these curves requires knowledge of water density and thickness of cloud. For further discussion of this refer to Chapter 2, Sections 2.1 and 4.2. Sky temperatures under cloud cover have been calculated and are plotted in Figure 39, using the model described in that figure.

4.1.3 Rain

Attenuation by rain is a complex process involving both absorption and scattering. For centimeter wavelengths and most drop sizes, absorption predominates and attenuation is approximately proportional to water content. However, millimeter wavelengths are comparable in size with the larger drops, so that scattering becomes important, and the frequency dependence of overall loss becomes complicated. Nevertheless, the problem has been solved, and the absorption and scattering coefficients of water drops are well established.[69] Thus attenuation can be calculated accurately for a specified density of drops of a specified size or size distribution. Hence the difficulty arises in knowing the drop density (or, equivalently, the integrated rate of rainfall along the path of the received signal) and the drop size distribution.

For horizontal paths along the surface of the earth, both of these quantities are, or can be, fairly well determined. Rate of rainfall traditionally has been measured with gauges which average over considerable periods. However, to evaluate the reliability of a microwave relay system, rainfall rates should be resolved in time down to seconds over a period of a year. Recently a high speed rain gauge[61] has been designed with an output adaptable for computer analysis. The gauge has measured rain rates which change by as much as a factor of 10 in seconds. Another method for obtaining rain rates with short time-resolution

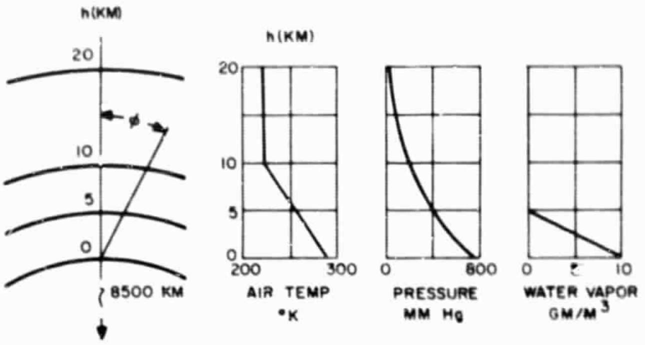


Figure 33. Model of the clear atmosphere

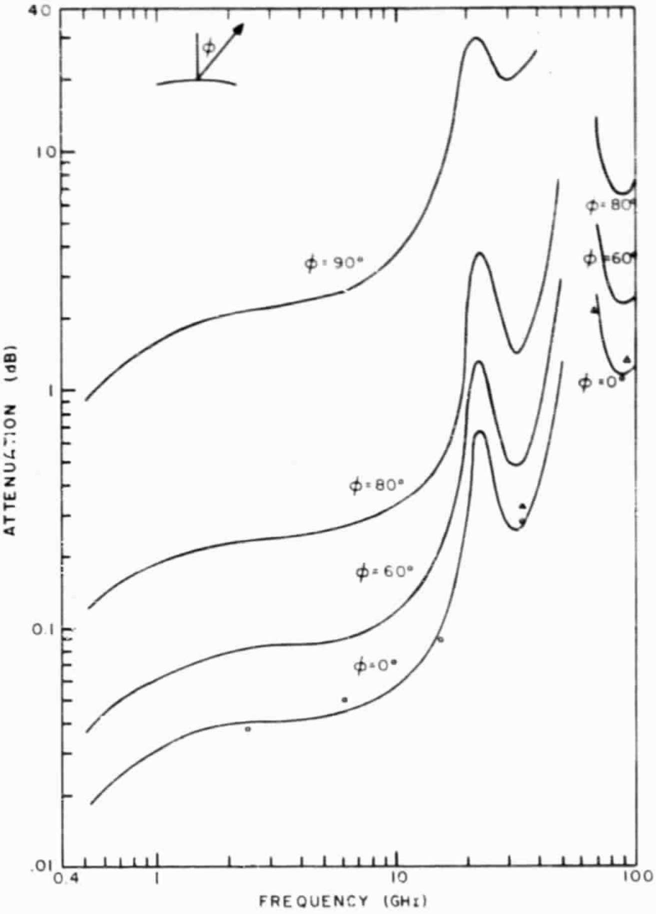


Figure 34. One-way attenuation through the total atmosphere (0_2 and H_20 vapor)

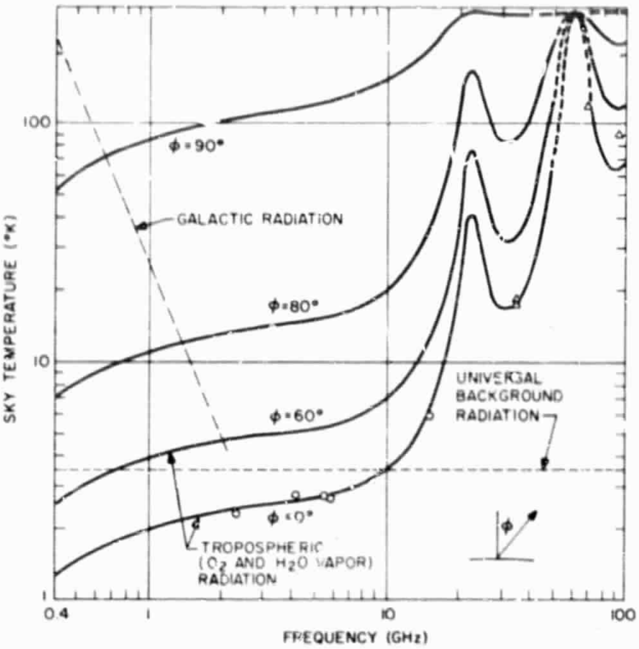


Figure 35. Sky temperatures (0_2 and H_20 vapor)

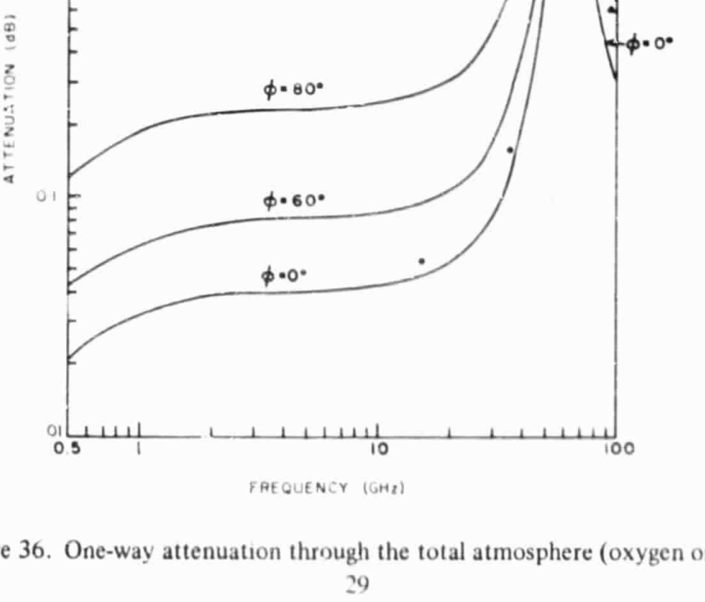


Figure 36. One-way attenuation through the total atmosphere (oxygen only)

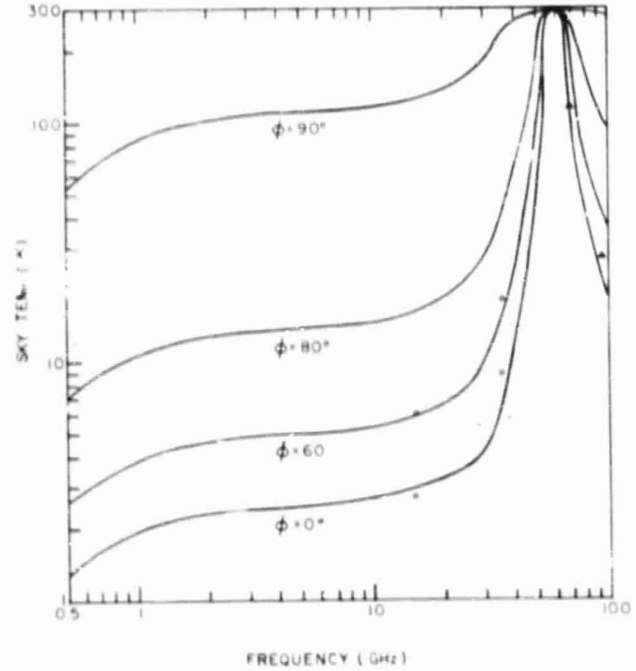


Figure 37. Sky temperatures (ox, gen only)

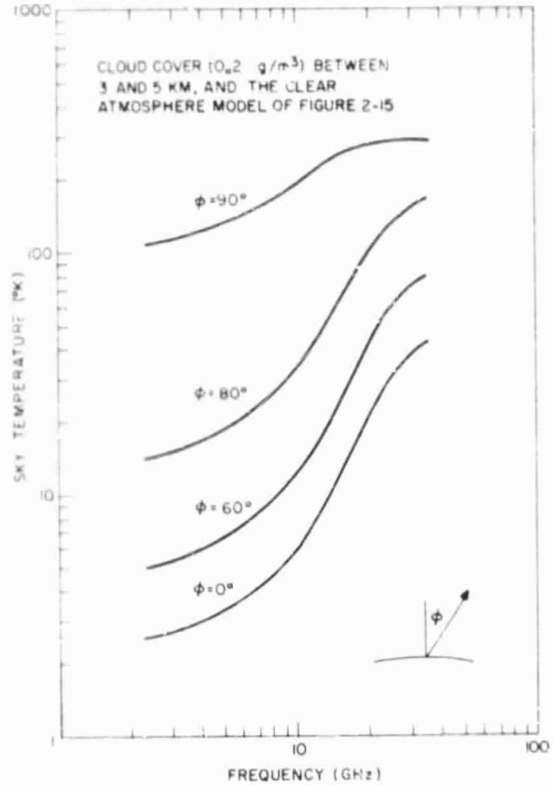


Figure 39. Sky temperatures under cloud cover

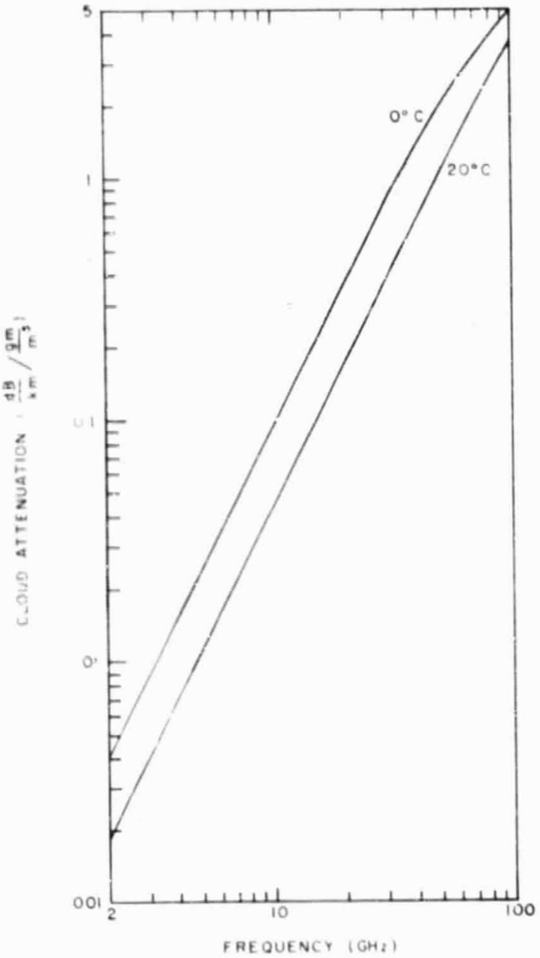


Figure 38. Cloud (fog) attenuation coefficients

was successfully developed by the Illinois State Water Survey.[62] Drops in a given volume were photographed, counted, and measured at various locations in the United States. The data are available[60] as percentage-time distributions for those locations.

The rate of rainfall measured at a point must be converted to a spatial average along a path to determine attenuation for an Earth-based radio relay system. This can be done readily if it is known that space-time ergodicity is valid: i.e., are the distributions of the time-average at one point and of the space-average along a path similar? This has been found to be the case[60] for data taken over four years on a line of four gauges with a spacing of 1 kilometer at Bedfordshire, England. To the extent that this may be generally true, it will be possible to use time-averaged distributions to predict average rates along horizontal paths.

Knowledge of the instantaneous spatial distribution of rainfall at the ground is required for planning route-diversity for Earth-based microwave relay systems. Such information is being obtained on two rain gauge networks. The network in Bedfordshire, England, is about 3 by 3 kilometers square with an intergauge spacing of about 1 kilometer. The other network, at Holmdel, New Jersey, is about 13 by 13 kilometers square with an intergauge spacing of 1.3 kilometers. This has 96 of the high-speed gauges[61] discussed above, and the time interval between successive maps of rain distribution can be as short as 10 seconds. These networks have provided quantitative confirmation[60] of what seems indicated from everyday experience: that rain, and especially heavy showers, have rather fine-grain spatial characteristics; i.e., the very large rainfall rates which would cause large attenuations tend to occur in cells of small lateral extent. As a result, reliability can be improved by a factor of 10 by switched-path diversity between two parallel paths 2 kilometers apart.[60]

Besides rain rate, the other quantity required for prediction of attenuation is the drop size distribution for the rain. Laws and Parsons[65] measured the distributions of drop sizes for rain at the ground. Their spectra are generally accepted and used in calculations of microwave attenuation by rain. Even so, there is not uniform agreement between measured attenuations and the thoretical predictions. The earliest such calculation is that of Ryde and Ryde.[63,64] Recently, Medhurst[71] recalculated their results for a wider range of parameters and pointed out some discrepancies. Medhurst concluded that available measurements do not entirely agree with theory; there is a tendency for measured attenuations to exceed the maximum possible levels predicted by theory. Blevis et al[72] made additional measurements at 8 and 15 GHz. They concluded that there is no definite tendency for measured attenuations to lie above the theoretical maximum values derived by Medhurst[71] and that the theoretical calculations provide a reasonable basis for the prediction of rain attenuation. Nevertheless, they found that the predicted values of attenuation are appreciably lower than those measured at low rain rates. Hogg, however, has obtained rather good agreement between measured and calculated attenuation by New Jersey rains at wavelengths of 6.2 mm and 4.3 mm.[60]

For prediction of sky noise-temperatures and attenuations appropriate to space communications, what is needed is similar information on rain rates (or integrated liquid water content) and drop size distributions for paths extending through the atmosphere at relatively large elevation angles. No such experimental data exist. Knowledge of the details of the structure of rain at the ground is incomplete; knowledge of the structure at higher altitudes is fragmentary at best. Weather radar measurements have shown that rain may originate as high as 45,000 feet.[73] The distribution of condensed water up to such an altitude would vary greatly with meteorological conditions, but the details of this are unknown. It has been proposed to use measurements by weather radar to calculate attenuation, but such a procedure would be unreliable. The radar return is proportional to the sixth power of drop diameter, but attenuation goes roughly with the third power. Also, radar responds to ice particles, which are not significant for attenuation.

There remains the possibility of measuring rainfall at the ground and relating this to attenuation and noise as observed by a microwave receiver which points up through the rain. Rather extensive measurements of zenith sky temperature at 6 GHz along with ground rain rate have been made by Hogg and Semplak.[73] Measurements were made during eight rain periods from March 3, 1960, to July 27, 1960. Several figures in their paper[73] show the variation with time of both zenith sky-temperature and ground rain rate. In some cases there is correlation involving a time lag, with variations in temperature preceding corresponding variations in rain. In other cases, such as shown by their figure for June 3, 1960, the noise measured was fairly high, but there was no measurable amount of ground rain. Both situations appear in the figure for June 18, 1960, which is here reproduced as Figure 40. In summary, "There appears to be no detailed correlation between measured ground rain rate and zenith sky temperature. . . ."[73]

It seems clear that large increases in sky temperature are associated with concentrations of liquid water within the field of the receiver, but the water need not appear as rain on the ground. Thus it is not possible to predict attenuations for paths through the atmosphere on the basis of ground rainfall rates. But it is possible to accumulate data on sky noise and attenuation associated with rainy conditions and to generate curves giving time distributions of noise and attenuation. Hogg and Semplak[73] have done this for the increase in zenith noise at 6 GHz during the eight periods of rain. Their figure is here reproduced as Figure

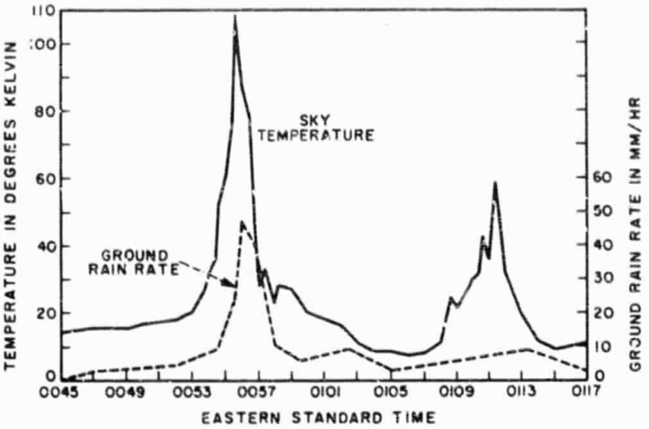


Figure 40. Example of lack of correlation between measured ground rain rate and zenith sky noise temperature (data of June 18, 1960 – from Hogg and Semplak, Reference 75)

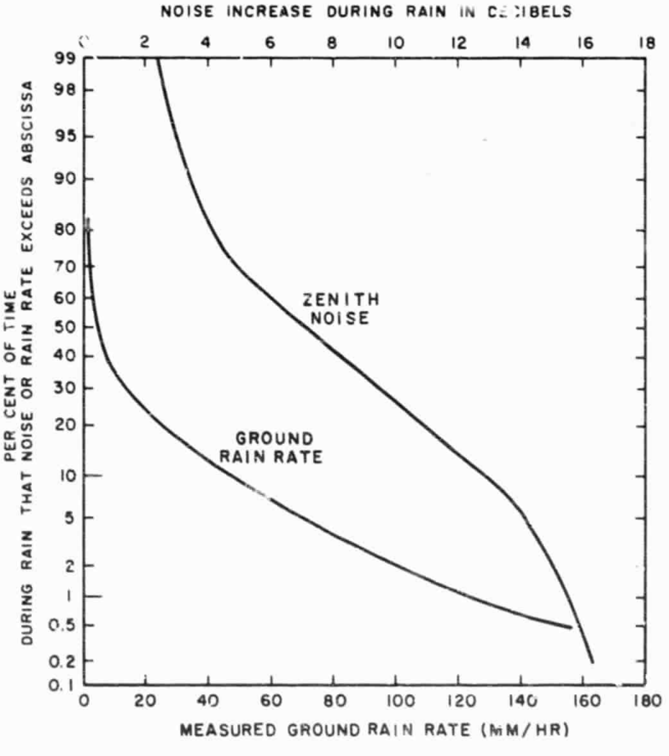


Figure 41. Percentage time distributions of zenith sky noise and ground rain rate, for eight rain periods from March 3 to July 27, 1960

41. Hogg and Semplak point out that this is a relatively small statistical sample and that additional measurements are needed, especially at beam positions nearer the horizon. It is also true that the noise distribution (Figure 41) does not necessarily apply to any other general location, even though the ground rain distribution for that location is known. Suggestions have been made that such distributions be extrapolated to other geographical areas by applying a correction factor such as the ratio of the mean annual precipitation between the two areas. This cannot be justified on the basis of present knowledge.

What can be done is to calculate noise and attenuation at other wavelengths using the best estimates for the dependence of attenuation on wavelength. Hogg has done this[60] to obtain a distribution of attenuation for 30 GHz using the above results (Figure 41) for 6 GHz. The result, given in Figure 42, applies to a zenith path above central New Jersey.

Wulfsberg measured the sky noise in the Boston area from February through July, 1963.[74] Whereas Hogg measured during rain periods, Wulfsberg accumulated data for all weather conditions. Noise was measured at 15 and 35 GHz and for zenith angles from 0 to 87.5 degrees. Wulfsberg pointed out that heavy rain was not encountered, so that more extensive measurements are needed for good statistics. Nevertheless, he provided more information of the type needed: a family of time distributions of noise at 15 GHz (Figure 43) and 35 GHz (Figure 44), with zenith angle as parameter.

More recently, Wulfsberg has reported measurements of attenuation at 15 and 35 GHz for zenith angles from 0 to 89 degrees.[75] Measurements were made daily over a six-month period using the Sun as a source. He obtained families of attenuation distributions for 15 GHz (Figure 45) and 35 GHz (Figure 46), with zenith angle as parameter. Wulfsberg's results are applicable only to the Boston area and perhaps to areas having comparable climates. Wulfsberg stated that extrapolation of the data to other geographic areas is difficult.

Gibble has obtained noise distributions at 5.35 GHz during rainfall in central New Jersey.[76]

Distributions of the kinds presented above are the basic data on noise and attenuation needed to plan space communications. Regardless of whatever correlation there may be with rain, such curves allow one to determine the probability that noise or attenuation will exceed a specified value for a specified elevation in a specified location. It is evident that more such data should be obtained, particularly at the existing (or prospective) locations for microwave space receiving stations. Less attention should be paid to rainfall statistics as such at these locations, because ground rainfall will not, at the present time, allow reliable predictions of noise and attenuation to be made for vertical paths through the atmosphere.

4.1.4 Path Diversity

The fact that heavy rain has a fine-grain horizontal structure[60] suggests that path diversity should substantially reduce the time that the link experiences a specified level of attenuation, just as it does for Earth-based relay systems. The primary question here is what separation is required between ground stations such that a heavy storm over one probably does not appear over the other. Weather-radar data taken over a five-month period at Montreal, Canada, were analyzed[77] to obtain the areas of storms of various intensities at various altitudes. Although storms come in assorted shapes, an average value for the suitable interstation spacing may be estimated by taking the square root of these areas. Such a set of equivalent diameters is given in Figure 47. The curves give the number of storms of a specified equivalent diameter whose intensity exceeded an equivalent rate of 25 mm/hour. This is done for various altitudes from 5000 to 40,000 feet. It is apparent that average diameter does not change much with altitude, so that good path diversity should result from an interstation spacing of 10 miles or more. Whether such a conclusion would be warranted for other geographical locations is not clear.

4.1.5 Snow and Ice

There is good general agreement that solid water produces negligible noise and attenuation at microwave frequencies. Noise measurements made at 6 GHz during a snowfall showed only a slight increase over those obtained on a clear day.[73] Wulfsberg stated that cirrus clouds composed of ice crystals produce a negligible contribution to sky noise at 15 and 35 GHz,[74] and that attenuation from cirrus clouds was not measurable at either frequency.[75]

4.2 Error in Prediction of Refraction

The microwave refractivity of the atmosphere may be expressed in terms of standard weather data as follows:[78]

$$N \equiv (n-1) \times 10^6 = \frac{77.6}{T}\left(P + \frac{4810\, e_s RH}{T}\right) \qquad (8)$$

where N is the index of refraction, T = temperature in degrees Kelvin, P = total atmospheric pressure in millibars, RH = percent relative humidity, and e_s = saturation vapor pressure in millibars. The atmospheric bending is independent of frequency in the microwave range because the refractivity is not a significant function of frequency here.

Atmospheric refraction is determined by ray tracing through the refractive index profile. However, the detailed

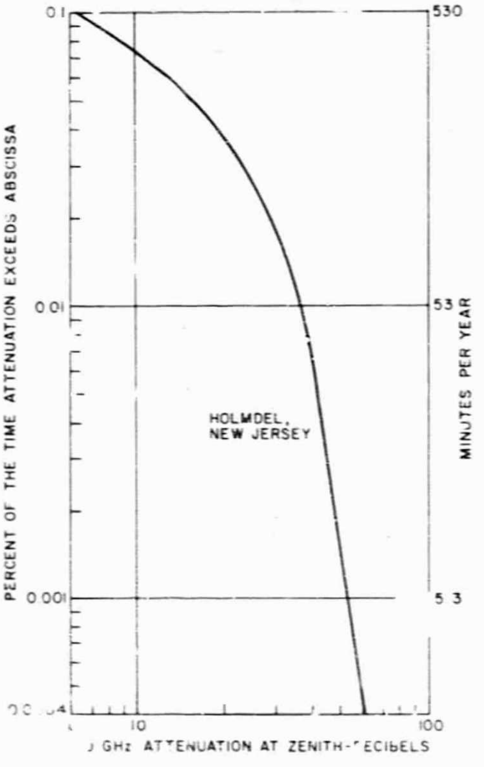


Figure 42. Estimated distribution of 30 GHz attenuation for a zenith oriented antenna

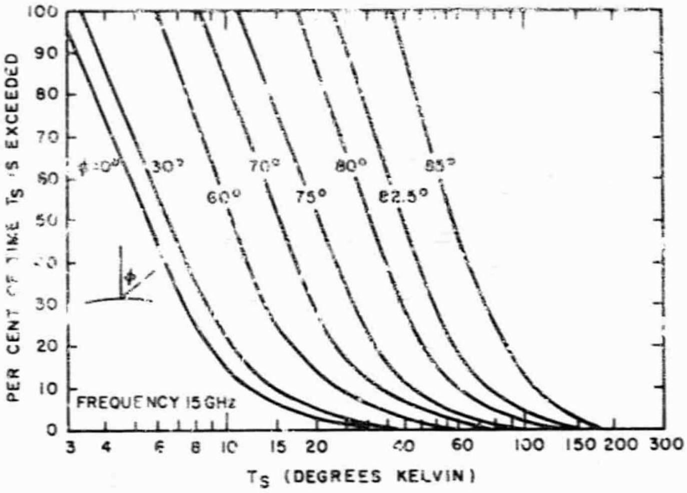


Figure 43. Sky temperature distributions, 15 GHz

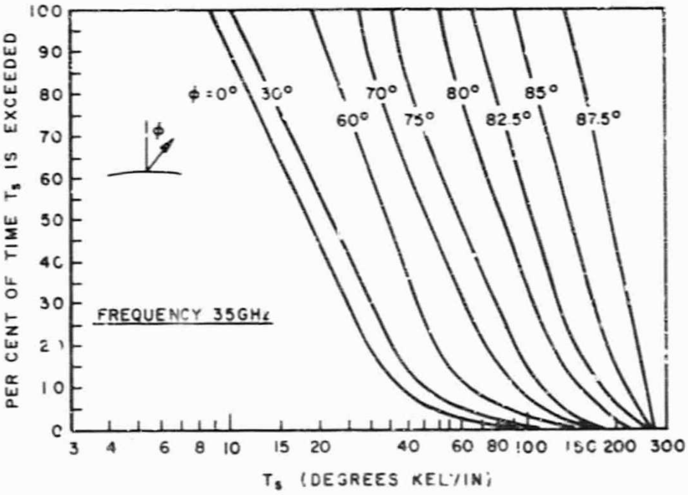


Figure 44. Sky temperature distributions, 35 GHz

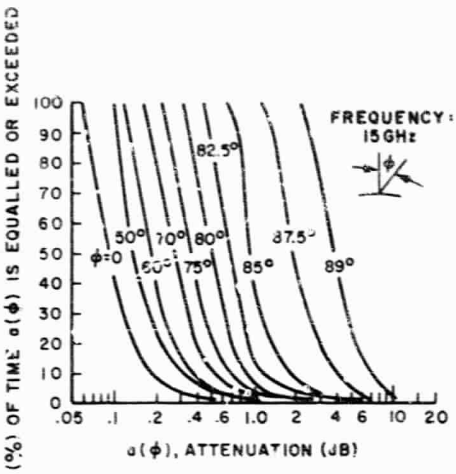


Figure 45. Attenuation distributions, 15 GHz

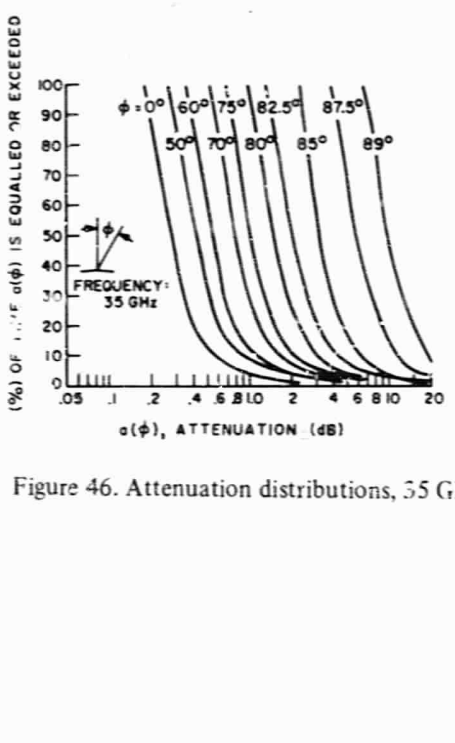


Figure 46. Attenuation distributions, 35 GHz

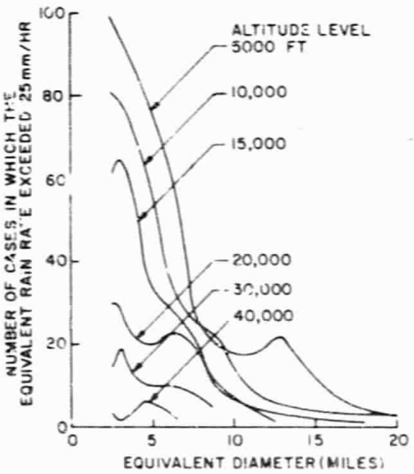


Figure 47. Frequency of occurrence of rain cells as a function of size for various altitudes

profile of the refractive index along the path of the radio ray is seldom available. Fortunately, it is possible to use the surface refractive index[79] to predict the total atmospheric bending of radio rays incoming from a deep space probe and arriving at a ground station.

The total bending τ (in radians) may be expressed as:

$$\tau = N_s 10^{-6} \cot\theta_o - \int_{(\cot\theta)_{N=0}}^{\cot\theta_o} N 10^{-6} d(\cot\theta) \qquad (9)$$

where N_s is the refractivity at the surface of the earth and θ_o is the elevation angle of arrival or departure of the ray at the surface of the earth. The second term of Equation (9) contributes less than 3.5 percent of the total for $\theta_o = 10$ degrees and becomes negligible as θ_o increases. Thus, for space communication conducted at elevation angles greater than 10 degrees, prediction by the first term of Equation (9) alone provides an accuracy of at least 10^{-4} rad.

For lower elevation angles, as small as 10 mrad, a linear regression equation is available for improved prediction. Below this the ray might become trapped in a ducting profile at which point prediction is very difficult. The regression equation is:

$$\tau = b\,N_s + a \qquad (10)$$

where the coefficients a and b have been obtained by Bean and Cahoon[79] from refractive index profile samples for a wide range of meteorological conditions at 13 climatically diverse U.S. radiosonde stations. The above prediction has been verified by many measurements on atmospheric radio refraction effects.[80]

A typical value of the surface refractivity is 313 for the worldwide standard atmosphere. The maximum elevation angle bias without correction will be about 4 millirads at 5 degrees elevation. With a correction based on the daily average of surface refractivity, the tropospheric bias at 5 degree elevation will be of the order of 0.1 millirad.

Observations of the Early Bird communication satellite at Andover, Maine, found the standard deviation of random tropospheric angle errors between 10 and 65 μrad.[81] This range of values compares quite well with other estimates.[82,83] Therefore, the total tropospheric angle error due to imperfect prediction of ray bending and random fluctuations of refractive index is expected to be less than 10^{-4} rad at an elevation angle of 30 degrees.

5. GROUND ANTENNAS

In this section ground antennas for both microwave and millimeter applications (frequencies from 1 to 94 GHz) are considered. Perspective would be lost, and an awkward treatment would result if the discussion were divided according to frequency.

5.1 Introduction

Mathematical models including five significant variables involved in ground antenna operation have been developed for both exposed and radome-sheltered structures. The five variables are diameter, cost, gain, frequency, and rms surface tolerance. Diameters of interest lie in the 10- to 250 ft range for exposed antennas and in the 30- to 500 ft range for antennas enclosed by a radome. Frequencies of interest vary from 1 to 100 GHz. Only conventional reflectors are included in this section. No consideration is given to actively controlled surfaces, multiple antenna synthetic apertures, or other such concepts which may be important in the future. Arrays are discussed in Appendix 2.

The first relation introduced in this study is Ruze's formula, which relates gain to diameter, frequency, and rms surface tolerance. In accepting Ruze's formula, one must also accept its shortcomings. The next two relations of interest, relating rms surface tolerance to diameter and cost to diameter, are deduced from information available on existing and proposed installations. In many cases, this information is vague and ambiguous. This situation has been dealt with by basing both the rms and the cost relationship on only three data points in the case of exposed antennas, and on four points (which represent the totality of available data) in the case of antennas with radomes. The points used are believed to be consistent and span the diameter range of interest. The result of this interpretation is a cost curve which represents the cost of an antenna with a certain rms surface tolerance. The specific correlation of cost, diameter, and rms surface tolerance is a novel feature of the present approach. Finally, a quality factor is introduced which associates departures from standard rms surface tolerance with departures from the standard cost curve. The functional relations chosen to relate these quantities were justified with arguments which are largely heuristic, since available data did not permit a more precise determination. However, when existing information on cost and rms surface tolerance was adjusted to a common standard using the functions chosen, the resulting agreement was gratifying.

The results of this exercise are expressed as two equations among the five variables of interest. A different set of two equations was obtained for each of the systems considered (exposed and enclosed). Although more complex than the simple power law relationship often suggested for the cost vs. diameter of ground antennas, this set of equations contains more information. Not only do the equations yield information about any specific case, but they also provide a starting point for various optimization studies. Three examples are given in this section; they deal with minimum cost gain-limited antennas, maximum cost-effective antennas, and minimum cost antennas for a specified gain and frequency. The last two cases were examined for both exposed and enclosed antennas. These examples also suggest others that could easily be done. The models are believed to be superior to others that have been suggested elsewhere.

5.2 The Nature of the Problem

Most of the effort was devoted to the determination of appropriate rms surface tolerance-diameter and cost-diameter relationships for ground antennas. A large number of reports were studied, and many personal contacts were made in the attempt to gather the significant and pertinent information. The functional relationships which emerge are the result of a distillation of these raw data. The confidence one is justified in placing in them is directly related to the data on which they are based. The distillation process unavoidably involved subjective judgement, but differences of opinion and contradictory results that come to light in comparing this study with similar studies carried out elsewhere will ultimately be resolved in terms of different interpretations of the raw data.

For a given antenna (or sample point), one needs to know the diameter, the cost, and the rms surface tolerance. There is seldom any need to question the reported diameter, but unfortunately the situation with regard to the other two quantities of interest is not as satisfactory. The problems involved in measuring the surface tolerance of a large paraboloidal reflector are themselves difficult, and measurement is an expensive and time-consuming task. User demands on high-performance antennas often are sufficient to prohibit the measurement of the rms tolerance in any sort of a statistically satisfactory way. When an existing structure is measured, the measurements are generally taken using surveying techniques or by some kind of mechanical device which traverses the surface. These techniques themselves place constraints on the structure which are sometimes unrealistic in terms of operational requirements (i.e., zero zenith angle, benign environmental conditions, etc.). Such data are then passed through a data-reduction process, the details of which are seldom disclosed, to determine the tolerance figure which is eventually reported. In one or two isolated instances, tolerance is determined by measuring gain over a range of frequencies and then using Ruze's gain equation to calculate rms surface tolerance. This would seem to be a powerful and effective technique,[54] but it supposes a knowledge of the aperture efficiency at each frequency – a quantity that is extremely difficult to determine independently.

There is a disturbing lack of consistency in the reporting of surface tolerance. It is generally measured with respect to the best-fit paraboloid, but it also can be referenced to the original design contour. It can be a deviation normal to the reflector surface or normal to the aperture plane. The distinction is seldom drawn. In some instances, maximum peak-to-peak deviations are reported. Ruze suggests that a factor of 3 can be used to convert such numbers to rms values.

Finally, the nature of the information depends on who supplies it. Antenna manufacturers normally have rms information of some sort, since they are usually required to demonstrate a specified tolerance. Rare is the manufacturer who will admit his product failed to comply, but manufacturers are also understandably loathe to disclose what contrivances or bending of definitions, if any, were necessary to meet required standards. The users, on the other hand, have a different point of view. In general, they are much less concerned with rms tolerances per se. If the device operates within 1 or 2 dB of expected levels, they are inclined not to quibble.

In short, it is generally not hard to obtain something called the rms surface tolerance of a given reflector, but it is exceedingly difficult to compare this number with similar numbers for other reflectors, or to relate it to any sort of a common standard.

For different reasons, the situation with respect to costs is even more vague. A high-performance antenna is a custom-made item. Hence its price must include appropriate R&D, engineering, tooling, and fabrication costs, which are difficult to determine with any precision and which cannot be distributed over a large number of units to minimize their net effect. No supplier will even hazard a guess at the cost of an antenna of diameter D with a tolerance ϵ without a funded proposal study beforehand. Since the circumstances and requirements of each situation must be considered separately, this conservatism is perhaps justified. In addition, there are relatively few companies in the business of building such structures, and competition is fierce. Their reluctance to make, and perhaps subsequently be embarassed by, off-the-cuff quotes is natural.

It would appear that establishing the cost of existing structures would be an easier matter. However, in the absence of any common standard, a tendency to modify the actual cost in the most advantageous direction is often observed. The degree to which this kind of bias influences reported costs cannot be determined, but it is not unusual to hear different stories from the buyer and the seller about the price of the same antenna.

The major items of uncertainty in determining costs is establishing exactly what the reported number of dollars bought. Here again the lack of a uniform standard is apparent. There are numerous ancillary items associated with a ground antenna that may or may not be included in the reported cost. These include electronics, feed structure, land acquisition, servo systems, data read-out, main and auxiliary power plant, support buildings, heating, lighting, and ventilation. The reported costs are seldom broken down in sufficient detail. Since some of the items mentioned above are very expensive, it is clear that a direct comparison of costs without a clear breakdown could be misleading. The alternative, to ignore questionable cost information entirely, is even less appealing.

At this point it is natural to ask what progress can be made, in view of all the treacherous areas that have been charted above. The answer is that considerable progress can be made, but with a couple of provisions. First of all, it must be kept in mind that t ie functional representations derived from the data and used to make up the mathematical model are the result of only the present interpretation. In spite of efforts to be objective, impartial, and 'horough, these representations contain a substantial dose of judgment and subjective opinion. Because of the present state of the information available, they cannot be otherwise. The results presented are clearly not unique and probably cannot even be called right or wrong, except in a qualitative sense. This is not the first study of this sort and certainly will not be the last. The results presented will certainly not agree in detail with others, since it is almost mandatory that each new interpretation of essentially the same data provides a somewhat different result, if only to justify the effort expended. The present study is no exception. Second, the mathematical model that is developed can be (and was) evaluated for answers with eight significant figures. To infer this degree of quantitative precision in actual practice would be nonsense.

5.3 Gain Formula

In 1952, Ruze suggested a formula for the gain of a reflecting antenna.[84] This formula has been generally accepted by antenna designers and communications engineers in spite of several restrictive assumptions incorporated in its derivation. These assumptio: s have been clearly restated by Ruze in his 1966 article.[51] Ruze's formula states:

$$G \cong \eta \left(\frac{\pi D}{\lambda}\right)^2 \exp - \left(\frac{4\pi\epsilon}{\lambda}\right)^2 \qquad (11)$$

Here D is the reflector diameter, λ is the wavelength at the frequency of interest, ϵ is the rms deviation of the reflector surface, and η is the aperture efficiency, a measure of the overall electronic properties of the antenna. D, λ, and ϵ must be in consistent units.

The first factor in Ruze's formula is the gain to be expected with a perfect, uniformly illuminated reflector. The effect of deviations from a perfect paraboloid are contained in the exponential factor, although no distinction is made between manufacturing inaccuracies and random deflections of the reflecting surface due to environment. The gain of a given antenna, with specified diameter and surface tolerance, increases as frequency is increased. However, a point is reached at which the exponential factor takes over, and a further increase in frequency results in a decrease of the gain. This point, at which the gain is a maximum for a given reflector, is called the gain-limit point.

Much the same type of behavior is noted if the operating frequency is held fixed and the diameter is varied. The cause for a "gain-limit" point in diameter is not immediately apparent from inspection of Equation (11), but it occurs simply because the rms surface tolerance is a function of diameter. Stack[85] has pointed this out in his work, and curves of gain vs. diameter for a number of frequencies can be found in his report.[86]

The aperture efficiency, η, includes the effect of nonuniform illumination, spillover, aperture blockage, front-end losses in antenna electronics, and other similar factors which contribute to degradation in performance. It specifically does not include the effects of an unperfect reflecting surface, at least as used in this report. For a well-engineered antenna, η should lie between 0.65 and 0.75, but this general statement provides no assurance that the aperture efficiency of a specific antenna is indeed within this range. Reliance on published information can also be misleading, since many authors tend not to define their particular concept of aperture efficiency carefully. This elusive quantity also depends to a certain extent on antenna geometry. The aperture efficiency of a Cassegrain antenna certainly differs from that of a focal point reflector, even for identical reflecting surfaces. A still different value would be found for an open Cassegrain antenna. Finally, the aperture efficiency is a function of operating frequency and the required noise temperature. However, no information is available on the specific character of such functional relationships.

As a result of the uncertainty associated with the aperture efficiency, it has been assumed that, when Equation (11) is used in this report, the aperture efficiency is taken to be 70 percent. Certainly, this represents a rather gross assumption, but no more sophisticated relation can presently be justified.

The gain calculated by Ruze's formula[51] is the gain "at antenna," so to speak. Atmospheric effects on signal strength such as turbulence or rain are specifically not included. These effects may be extremely important, particularly at high frequencies, but are not explicitly part of the ground antenna considerations.

5.4 Relation Between Surface Tolerance and Diameter

The data available on rms surface tolerance of existing antennas are plotted in Figure 48. The ranges associated with many of the data points are attributable to a number of factors. In a few cases they reflect an honest uncertainty. In

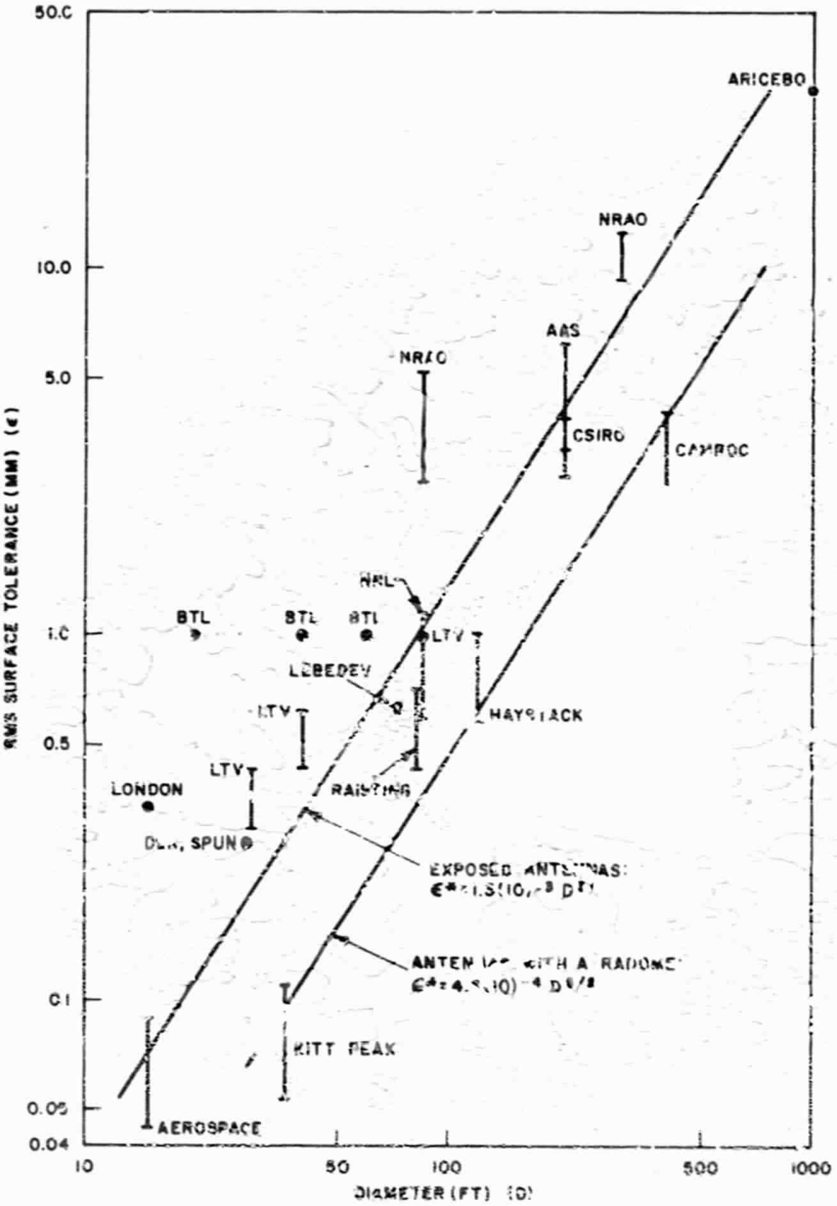


Figure 48. Diameter vs. surface tolerance for existing antennas

others they are due to different tolerances reported at different elevation angles, or under different environmental conditions. In some cases the range shown is a result of different values reported from different sources for the same antenna. For one or two antennas, the range shown represents the design goal and the performance evidence has not yet been reported.

The surface tolerance data represent the surface accuracy under operating conditions. Thus, all factors that combine to produce mechanical deviations from a perfect paraboloid are included. These include manufacturing inaccuracies as well as surface deflections caused by environmental loads such as gravity, wind, and thermal effects. In general, the antennas represented are fully steerable and operate satisfactorily in steady winds up to about 30 mph or so and in other environmental conditions normally expected for such antennas.

In spite of the considerable scatter of the data on Figure 48, it is reasonable to represent general ϵ -D trends by two straight lines: one for exposed antennas, and the other for antennas operated inside a radome. The appropriate functional relationship is of the form:

$$\epsilon^* = a D^{3/2} \qquad (12)$$

where ϵ^* is the rms surface tolerance in millimeters, and D is the reflector diameter in feet. The reader is cautioned to take special note of this rather unusual juxtaposition of units. The constant a is:

$a = 1.3(10)^{-3}$ for exposed antennas

$a = 4.6(10)^{-4}$ for antennas under a radome.

More must be said about the rms-diameter relation for exposed antennas, particularly in view of the scatter of the data for antennas of this type. However, further discussion is deferred until the section on cost-diameter relations, where it can be included more naturally. The curve for reflectors under a radome is based on regrettably few points (four, to be specific), but additional data are not available. The curve has the same slope as the one for exposed structures, but, at any given diameter, the surface errors are considerably less for the enclosed structure because of the benign environment.

5.5 Relation Between Cost and Diameter

A plot of cost vs. diameter for ground antennas contains a scatter of the data at least an order of magnitude worse than the data presented in Figure 48. A straight line "fit" (on the standard log-log plot, which corresponds to the power law relation Cost = (Const.) D^n, is simply unacceptable over the entire size range of interest. A piecewise-linear cost function, corresponding to an increase in the power law exponent with diameter, would be much better but would introduce troublesome analytic complications.

The uncertainty introduced by this kind of visual curve fitting is appalling, but the method is probably no worse than attempting to establish an rms best-fit curve in logarithmic coordinates. In lieu of such methods, three basic antenna structures have been chosen. The points span the size range of interest and can be fit by a three-parameter expression. The three antennas chosen are:

1. The 15-foot antenna operated by Aerospace Corporation, El Segundo, California
2. The 85-foot antenna operated by the Naval Research Laboratory at Maryland Point, Maryland
3. The 210-foot AAS antenna operated by Jet Propulsion Laboratory at Goldstone, California.

All three antennas were manufactured by the same company; all are exposed and fully steerable, although the first two are polar mounts while the third is az - el; and reasonably good rms surface tolerance information is available for all three. More important, the cost data obtained from user and manufacturer agree to within 10 percent for the first two and agree exactly for the 210-foot antenna. The costs cited include structure, drives, and control but do not include electronics, readout equipment, or other ancillary costs, insofar as could be determined. The cost-diameter relation obtained by this process is

$$\$^* = 6.7(10)^5 D^{-1/3} \exp(D/45) \qquad (13a)$$

This curve is shown in Figure 49†. In Equation (13a), the diameter D is in feet. Potter's power law curve[87] for the 85 to 250 ft range is also shown in Figure 49.

Although Equation (13a) fits the three selected points very nicely, problems occur if unconscious extrapolation is attempted. Beyond 210 feet, the costs increase rapidly with diameter because of the exponential factor. On the other end there is a singularity at D = 0, and costs again increase as diameter decreases below 15 feet. Upon reflection, neither of these tendencies is entirely unrealistic. However, it is recommended that Equation (13a) be used only over a diameter range of 10 to 250 feet.

The cost-diameter relation for antennas with a radome is also shown in Figure 49. In this case, the items included in and excluded from the reported cost are the same as for the exposed antennas with one important exception—the cost of the radome is included. A power-law relation is satisfactory for these antennas over a diameter range of 30 to 500 feet. This relation is

$$\$^* = 6.75(10)^3 D^{1.30} \qquad (13b)$$

Again, the diameter is expressed in feet.

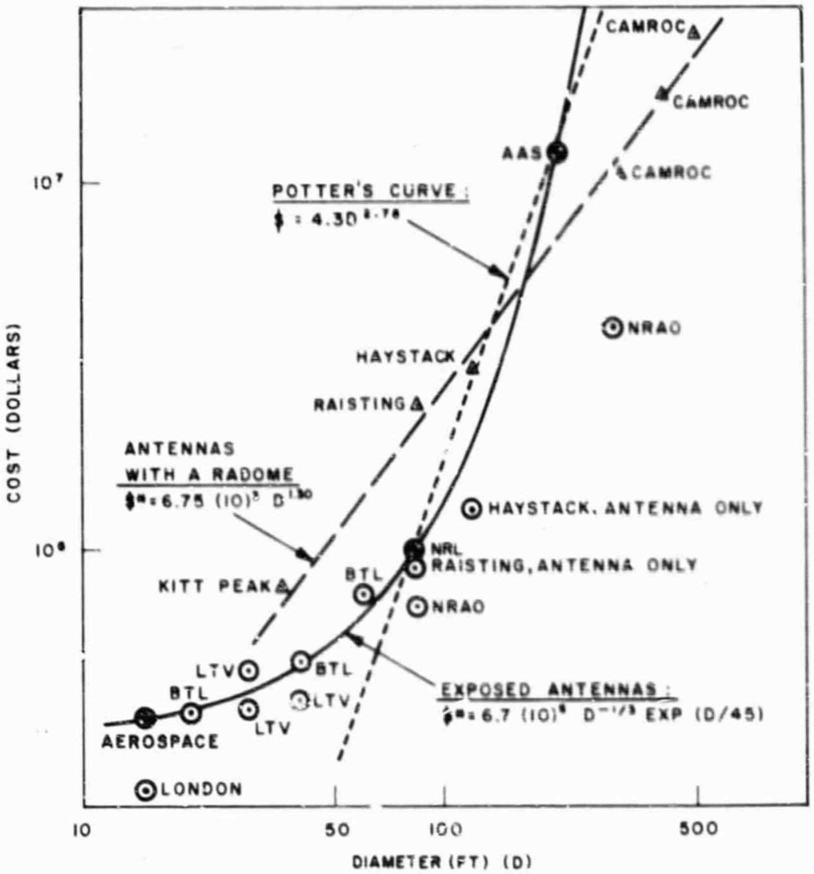


Figure 49. Diameter vs. cost for existing antennas

†The data points scattered about the curve on Figure 49 do not represent raw data for other antennas. The significance of these points will be discussed in the next section.

Experience with the operation of radome-enclosed antennas has been satisfactory at microwave frequencies. The radome is responsible for approximately 1- to 1.5-dB loss in gain, primarily because of aperture blockage, and it also contributes to system noise temperature. The system degradation depends to a considerable extent on local weather conditions. A discussion of these effects is beyond the scope of this report, but a thorough explanation of the electronic properties of radomes can be found in the CAMROC report.[88] There is little experience with radomes at millimeter frequencies. In this region the radome thickness is no longer small with respect to wavelength, and special design techniques would clearly be necessary to minimize losses.

The cost of the radome alone, including foundations and environmental control equipment, is shown in Figure 50. Both air-supported and rigid space frame radomes are included. This information is taken directly from the CAMROC report[88] for the rigid radomes and from the BTL antenna study[89] for the air-supported radomes. A gross extrapolation is required in each case to include the entire diameter range of interest. In lieu of an acceptable alternative the required extrapolation is made, but not without considerable reservation. The diameter of the radome required to enclose an antenna of diameter D is assumed to be 4/3 D. The cost of the antenna alone can be determined by using Figures 49 and 50.

The antennas used as the basis for the cost curves also determine the rms tolerance-diameter curves (refer to Figure 48). Thus the cost-diameter relations [Equations (13)] give not just the cost of an antenna of diameter D, but specifically the cost of an antenna of diameter D with a surface tolerance given by Equation (12). This correlation between the cost-diameter and rms-diameter curves is extremely important. The two relationships, taken together, express cost in terms of diameter and rms tolerance. The step from rms tolerance to frequency is simple; thus, the cost is related to diameter and frequency, albeit implicitly. However, the frequency dependence of the cost function is identified by this approach. The two sets of curves, as given on Figures 48 and 49, must be interpreted together, not separately. The costs and rms surface tolerances defined by these curves will be referred to as standard rms and standard cost, identified by ϵ^* and $\$^*$.

Several well-known antennas have not been included in these plots, generally because of lack of data on either cost or rms surface tolerance. None of the good Russian antennas appear because no reliable cost information is available. Other antennas were excluded, even though all the information was available, if there was an obvious anomaly for which a reasonable explanation could be given. For example, the 140 ft dish at Greenbank, West Virginia, has an astronomical price tag because of faults in the original design. The 210 ft CSIRO antenna at Parkes, Australia, has a phenomenally low cost (1/6 of AAS) which is attributed to low labor costs and to the limited steering capability of the antenna. Such points are not included.

5.6 Quality Factor

The standard case is represented by Figures 48 and 49. The effects of moving off the given curves must now be examined. In other words, how much can be saved by relaxing the rms requirement at a given diameter, or conversely, how much will it cost to improve the surface tolerance at some given diameter? These questions lead naturally to a host of others, such as: Is it better to increase diameter or surface tolerance to achieve desired performance? From such questions the possibilities of a full-fledged optimization study begin to emerge.

To deal with such questions, the quality factor is introduced. In essence, it relates a change in rms surface tolerance to a change in cost. This ingenious approach to the problem was first introduced by Stack.[85,86] The actual RMS (ϵ) and actual cost ($\$$) are expressed as

$$\epsilon = f_1 \epsilon^*$$

$$\$ = f_2 \$^*$$

where ϵ^* and $\$^*$ are the standard values found from Figures 48 and 49. In the radome case the cost appearing in these relations must be the cost of the antenna alone. The problem reduces to a determination of the functions f_1 and f_2.

It is unnecessary to trace in detail the tortuous path leading to the selection of the functions f_1 and f_2. The functions ultimately selected have the form

$$f_1 = 1/\chi \qquad (14a)$$

$$f_2 = \exp(\chi - 1) \qquad (14b)$$

The quantity χ is defined as the quality factor. For $\chi > 1$, the rms surface error is less than the standard rms error given by Equation (12). Conversely, for $\chi < 1$ the surface is less precise than given by the standard curves.†

The range of the variable χ is $0 \leqslant \chi \leqslant \infty$. This range therefore includes the possibility of achieving a nearly perfect reflecting surface by requiring that χ be very large. Physically, of course, this is not possible. There exists a

†Stack[86] proposed the relations $f_1 = 1/b$, $f_2 = b$. He also imposed the constraint that $f_2 \geqslant 1$ for all b. Equation (14a) corresponds to Stack's original choice, but Equation (14b) removes the constraint Stack imposed on f_2.

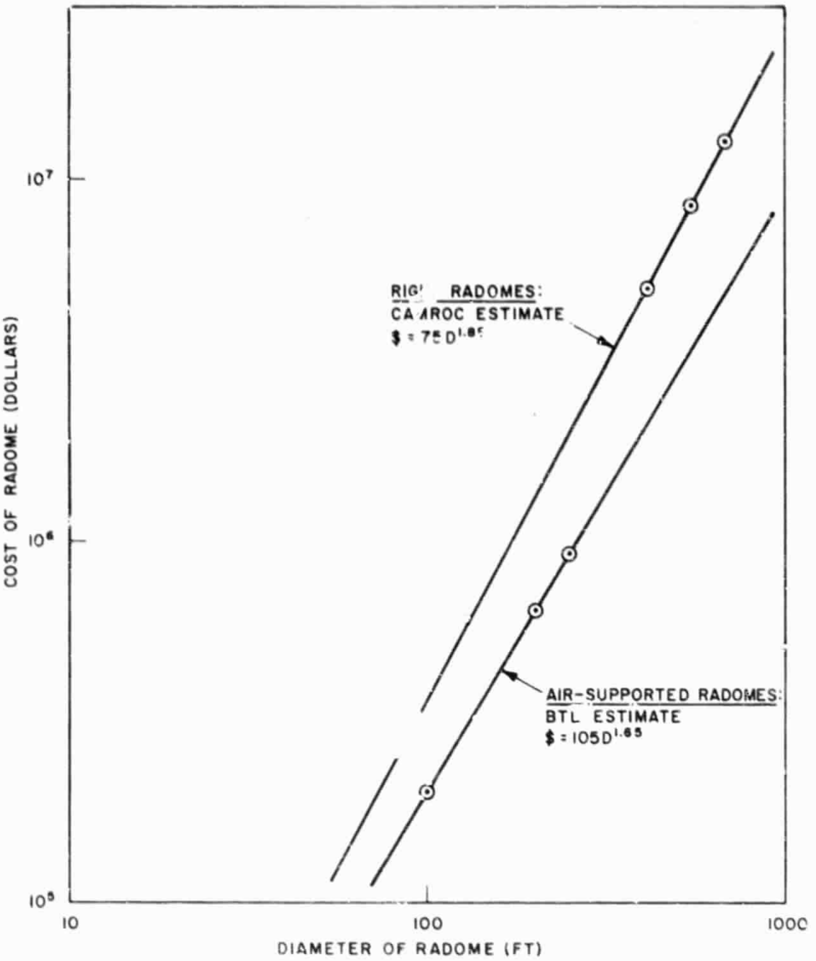


Figure 50. Diameter of radome vs. cost for radomes for large antennas

limiting tolerance, almost certainly a function of diameter, beyond which the surface accuracy can no longer be improved. Unfortunately, no one seems to know what that limit is. Equations (14) represent a compromise with this situation. Although Equation (14a) admits the possibility of infinite improvement in rms surface tolerance, Equation (14b) associates an infinite cost with such an improvement. In fact, the cost factor f_2 expressed by Equation (14b) extracts a very heavy cost penalty for even modest rms surface tolerance improvements. In addition, Equation (14b) limits the possible reduction of cost to approximately 1/3 of standard, regardless of the reduction of quality of the reflecting surface.

These two modifications in the properties of the cost factor are the major differences between this and Stack's original work.[86] They are based on the realization that the standard curves of Figures 48 and 49 represent very good reflecting surfaces. It is reasonable to expect that further improvement of the surface quality will be extremely expensive, while some saving should result if the standards of accuracy are relaxed. The cost factor expression, Equation (14b), would probably not be applicable if the three basic antennas from which the standard cost curve [Equation (13a)] is derived has been nearer the center of the spectrum of available products. However, the three points actually used represent a definite bias toward the excellent, and this bias justifies the form of the cost factor.

There is not enough good raw data on the rms tolerance and cost of existing antennas to establish the nature of the quality factor functions. To set up the quality factors, at least two relatively high confidence data points would be necessary for several different diameters. With the present standards of reporting, it was difficult enough to establish one such point. The quality factor functions given in Equations (14) should also somehow reflect the influence of diameter. It seems reasonable that it would be more difficult to achieve a given improvement in a large reflector than a small one. Including this effect in Equations (14) could not be justified on the basis of the present data.

However, the chosen quality factor functions were checked against the available raw data. For each antenna, the quality factor was determined by comparing actual and standard rms values according to Figure 48. The appropriate cost factor could then be found from Equation (14b). The inverse of the calculated cost factor was then applied to the reported cost of the antenna to find a revised standard cost. This represents the expected cost of that antenna if it had been built to the standard rms surface tolerance. The points obtained by performing this exercise for as many antennas as possible (i.e., those for which reasonable values for both rms tolerance and cost were available) are shown in Figure 49. While agreement is not perfect, it is at least considerably better than any possible fit to the unmodified raw data.

It should not be inferred that the quality factor functions of Equations (14) are unique in any sense. Other functions provide the same sort of qualitative trends. It is possible to suggest functions which include provision for placing finite limits on the possible improvement or degradation of the standard surface tolerance, and finite limits for the associated cost factor as well. These functions are somewhat more complicated than those actually chosen, and the implications of using them have not been investigated. Present information simply does not permit a definite choice to be made among all the possible quality factor functions that can be suggested. Hence, the compromise was to accept the set of functions given in Equations (14), which were both qualitatively reasonable and analytically convenient.

5.7 Mathematical Models

By combining the equations developed in preceding sections, one can find the equations appropriate for the study of the interrelations of gain, cost, diameter, surface tolerance, and frequency for both exposed and radome-enclosed antennas. For exposed antennas:

$$\epsilon = \frac{a_1 D^{3/2}}{\chi} \tag{15a}$$

$$S = a_2 D^{-1/3} \exp(a_3 D + \chi - 1) \tag{15b}$$

$$G = \eta (a_4 D \Omega)^2 \exp - (a_5 \epsilon \Omega)^2 \tag{15c}$$

Here, ϵ is the rms surface tolerance in millimeters, D the antenna diameter in feet, G the gain in absolute units, S the cost in dollars, η the aperture efficiency, Ω the frequency in GHz and χ, nondimensional, the quality factor. Appropriate values for the constants are:

$a_1 = 1.3 \times 10^{-3}$ $\quad a_2 = 6.7 \times 10^{5}$ $\quad a_3 = 2.22 \times 10^{-2}$

$a_4 = 3.20$ $\quad a_5 = 4.19 \times 10^{-2}$ $\quad \eta = 0.70$

Equations (15) are appropriate for a diameter range of approximately 10 to 250 feet.

For antennas in a radome:

$$\epsilon = \frac{\beta_1 D^{3/2}}{\chi} \qquad (16a)$$

$$\$ = \exp(\chi - 1)\,[\beta_2 D^{\beta_3} - \beta_4 D^{\beta_5}] + \beta_4 D^{\beta_5} \qquad (16b)$$

$$G = \frac{1}{1.26}\,\eta\,(\beta_6 D\Omega)^2 \exp - (\beta_7 \epsilon \Omega)^2 \qquad (16c)$$

The terms have the same meaning as for exposed antennas. The cost factor is applied only to the cost of the antenna, since the cost of the radome is obviously independent of the quality of the antenna inside. The total cost includes the cost of the radome. The expression for gain has been modified by a factor of 1.26, which corresponds to an assumption of a 1 dB loss caused by the presence of the radome. The system effect of the noise temperature added by the presence of the radome has not been considered. Appropriate values for the constants in Equations (16) are:

$\beta_1 = 4.6 \times 10^{-4}$	$\beta_2 = 6.75 \times 10^3$	$\beta_3 = 1.30$
$\beta_6 = 3.20$	$\beta_7 = 4.19 \times 10^{-2}$	$\eta = 0.70$

For rigid radomes: $\beta_4 = 1.28 \times 10^2$; $\beta_5 = 1.85$

For air-supported radomes: $\beta_4 = 1.69 \times 10^2$; $\beta_5 = 1.65$

Equations (16) are appropriate for a diameter range of approximately 30 to 500 feet.

5.8 Examples

Much information can be obtained from the models expressed by Equations (15) and (16). Each set is comprised of two expressions among five variables. Thus any three can be specified and the other two found directly. There are so many possible combinations that no general solution curves are given. It is simpler to enter the appropriate equations for each specific case and work out the result.

Of more interest are the various types of optimization studies that can be carried out using Equations (15) and (16). The results of three specific studies are given here. These are obviously not the only such studies that can be carried out. The details of the necessary algebraic manipulations are omitted, since they are generally straightforward but tedious. However, the procedure is indicated in each case. The discussion is presented in terms of Equations (15) for exposed antennas, although the procedure described is also appropriate for Equations (16), with obvious changes. For all examples including a radome, a rigid radome was assumed.

Example 1. Minimum-Cost Gain-Limited Antenna. In this example it is assumed that the antenna operates at the point of maximum gain; that is, the gain-limit point. At this point, the relationship between rms surface tolerance and frequency is[51]

$$\epsilon = (a_5 \Omega)^{-1}$$

When this relation is substituted in Equation (15a), it can be solved for $\chi(\Omega, D)$. By inserting this result in Equation (15b) the cost can be expressed as a function of D and Ω. The extrema of this function are found by equating its derivative with respect to D to zero and searching for roots. Such roots, if they exist, must be tested for maximum or minimum properties. The values of the roots D will depend on frequency.

The results of this example show that this cost, considered as a function of diameter and frequency, has no internal minima within its range of validity. There is a relative minimum at the lower end of the diameter range. Thus, for a minimum cost gain-limited antenna, one should use the smallest possible reflector, operate at the highest possible frequency consistent with other constraints on the system, and build the reflector to a surface tolerance corresponding to the gain-limit point at the operating frequency. Any increase in diameter will increase the cost but will also increase the gain.

This example was not carried out for antennas with a radome.

Example 2. Maximum Cost-Effective Antenna. A maximum cost-effective antenna is defined as one which provides the most gain per dollar of cost. As in example 1, it is assumed that the antenna operates at the gain-limit point, and the cost of the antenna is expressed in terms of diameter and frequency, as before. At the gain-limit, gain Equation (15a) becomes

$$G = \frac{1}{e}\,\eta\,(a_4 D\Omega)^2$$

This expression is divided by the cost expression to form the ratio G/\$. The maxima of this expression, considered as a function of D, are sought by using standard techniques. This time the search is fruitful. The expression G/\$ has a single maximum in the diameter range of interest. The location of the maximum depends on the frequency, as expected. The results of this example are plotted in Figures 51 and 52 for exposed and radome-enclosed antennas, respectively. The ordinates are diameter and cost, plotted against a common abcissa, frequency. The gain at the point of maximum cost effectiveness is cross-plotted along the diameter curve. There is a much greater range of gain with frequency for the exposed antennas than for antennas with

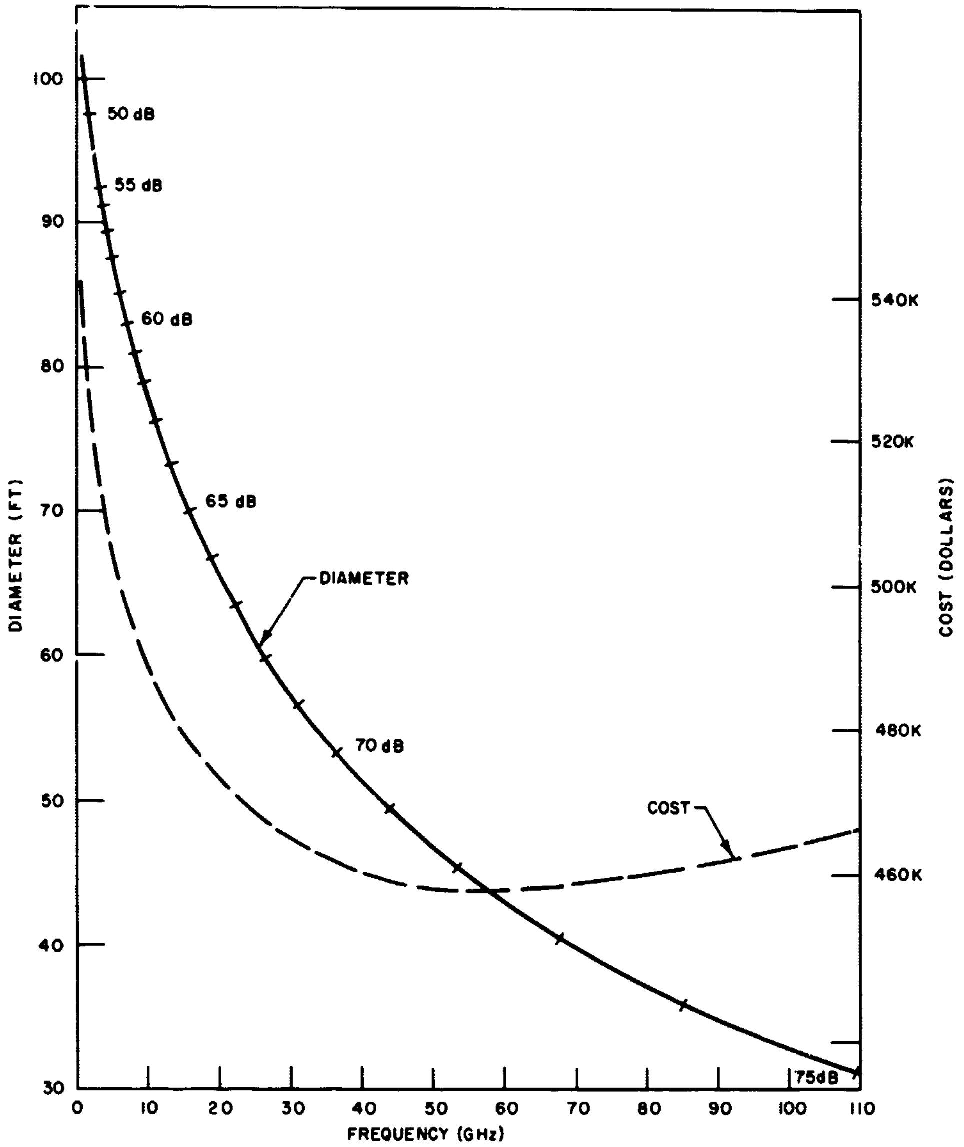


Figure 51. Maximum cost effective antennas (no radome), diameter vs. frequency

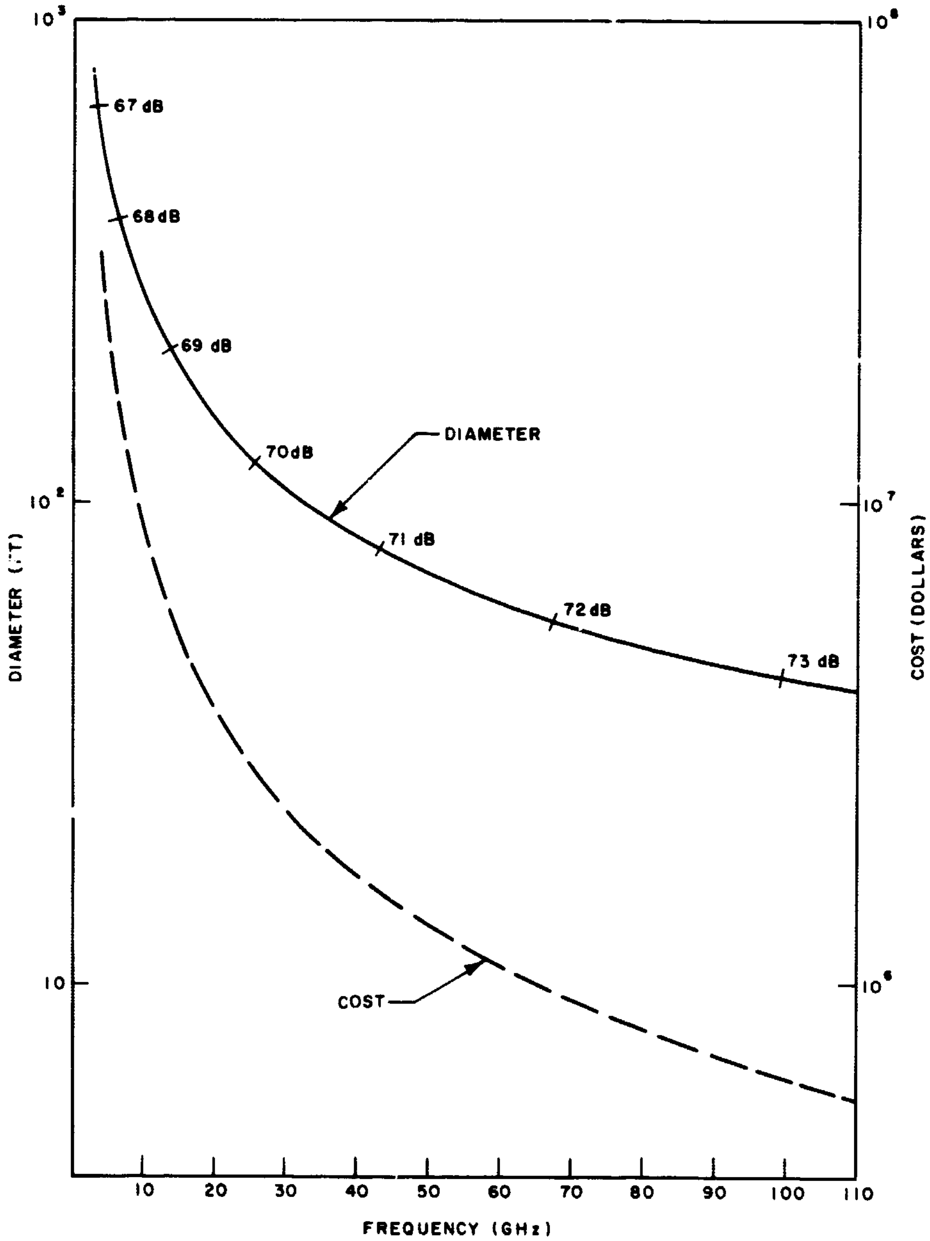


Figure 52. Maximum cost effective antennas (enclosed by radome), diameter vs. frequency

radomes. This is directly attributable to the diameters involved. No exposed antenna is larger than 105 feet, while, at the low frequencies, radome-enclosed antennas several hundred feet in diameter are found. The small variation in the cost of exposed antennas for all frequencies is surprising. All exposed maximum cost-effective antennas cost $500.000 ±10 percent, regardless of the operating frequency. The minimum in the cost curve at ~50 GHz also reveals the min-max properties of this example.

Example 3. Minimum-Cost Antenna for a Specified Gain and Frequency. In this example the operating frequency and the gain required of the ground antenna are specified in advance, perhaps as a consequence of other system constraints. The first step of the optimization procedure is to substitute Equation (15a) into Equation (15c). The resulting expression is then solved for the quality factor χ (D). The variables G and Ω have been fixed at their specified values. This expression for χ is inserted in Equation (15b), yielding cost as a function of diameter. The diameter that minimizes the cost is then found by differentiation of this expression. The algebra involved in this example is unpleasantly heavy, and numerical search techniques were used to determine the minimum cost diameter for both the exposed and the radome-enclosed case. The results appear in Figures 53 and 54 respectively. Figure 53a shows diameter vs. frequency and 53b gives cost vs. frequency for several different values of gain. The same pattern is found in Figure 54a and b.

For radome-enclosed antennas, the results of this example for diameter vs. frequency and cost vs. frequency are represented by a set of straight lines in log-log coordinates (see Figures 54a and 54b). Therefore, the solutions have the form $D = \phi(G)\Omega^n$ and $S = \psi(G)\Omega^m$, where ϕ and ψ are some functions of the gain only.

One variable (either gain or frequency) could be eliminated leading to a single expression for the cost in terms of diameter and the other remaining variable. However, for any specific case it is just as easy to use the curves given.

The diameter vs. frequency relations for exposed antennas are also straight lines in log-log coordinates for the lower frequencies, but exhibit a definite curvature at higher frequencies, particularly for the lower gains (see Figure 53a). The cost vs. frequency curves (Figure 53b) are not even approximately linear. They illustrate the exceptionally high cost of gain at low frequencies which results from the large-diameter antennas required.

An interesting comparison can be made between the results given by Figures 53b and 54b. For a specified gain, there is a range of frequencies (or diameters according to Figure 53a or 54a) in which an exposed antenna is less expensive than one enclosed in a radome. This range of diameters varies with gain. For 55-dB gain, an exposed antenna with diameter between 24 and 190 feet (corresponding to a frequency range of 1.18 to 11.6 GHz) is less expensive than an equivalent system with a radome. At 75 dB the range is reduced to 44 to 68 feet (37 to 60 GHz). For diameters, or frequencies, outside these limits the radome-enclosed system represents a better buy for a specified performance.

5.9 Summary

Mathematical models relating cost, diameter, gain, rms surface tolerance, and frequency have been developed for both exposed antennas and antennas in a radome. The form of the model in each case is a set of two equations among the five variables of interest. This model is more complicated than other cost vs. diameter relations suggested in the past, but it is considerably more general and can be used to study a variety of possible trade-off situations.

The major features of the models presented are:

1. The inclusion of an exponential factor in the cost vs. diameter relation for exposed antennas. This reflects, at least qualitatively, the exceptionally high cost associated with large, high-performance, exposed antennas.
2. The specific correlation of the cost vs. diameter and rms surface tolerance vs. diameter relations. As a result, costs are associated not only with diameter but also with rms surface tolerance.
3. The introduction of the quality factor. This factor relates changes in rms surface tolerance specification to expected changes in cost. Although qualitative in nature, this factor reflects acknowledged trends.

The two equations constituting the model will provide values for any two of the variables of interest, once the other three have been chosen. Additional relations among the variables, such as an rms surface tolerance-wavelength relation, for example, can be added. There is virtually no limit to the kinds of optimization studies and trade-off investigations that can be carried out within the framework of the suggested models. Examples of three such studies have been included. Through the credibility of the results, these examples further demonstrate the qualitative validity of the models. Minor revisions in the constants of the present model, resulting from new information or even from different interpretations of present data, are to be expected and encouraged. However, such refinements should not invalidate the general applicability of the present model nor any qualitative conclusions drawn from its use.

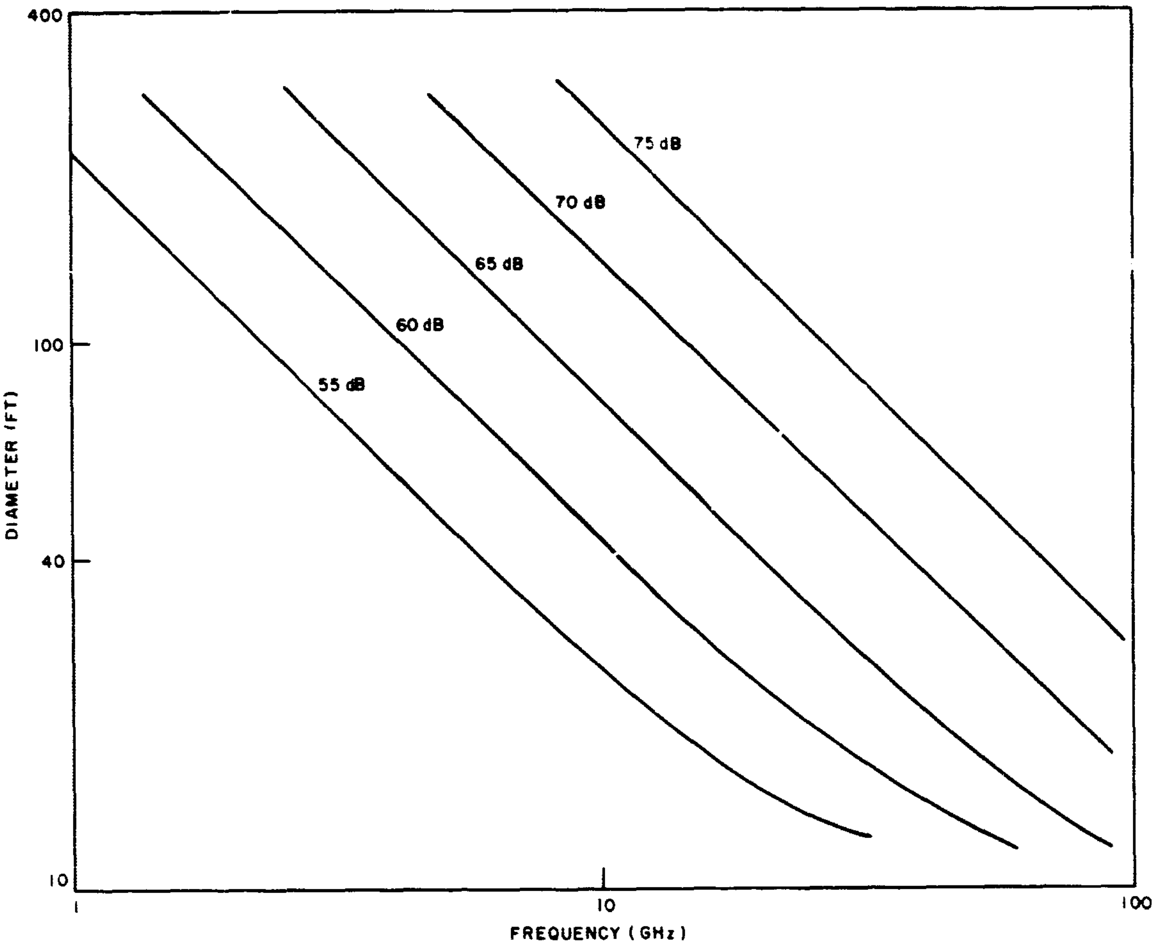


Figure 53a. Minimum cost antenna: fixed gain and frequency (no radome), diameter vs. frequency

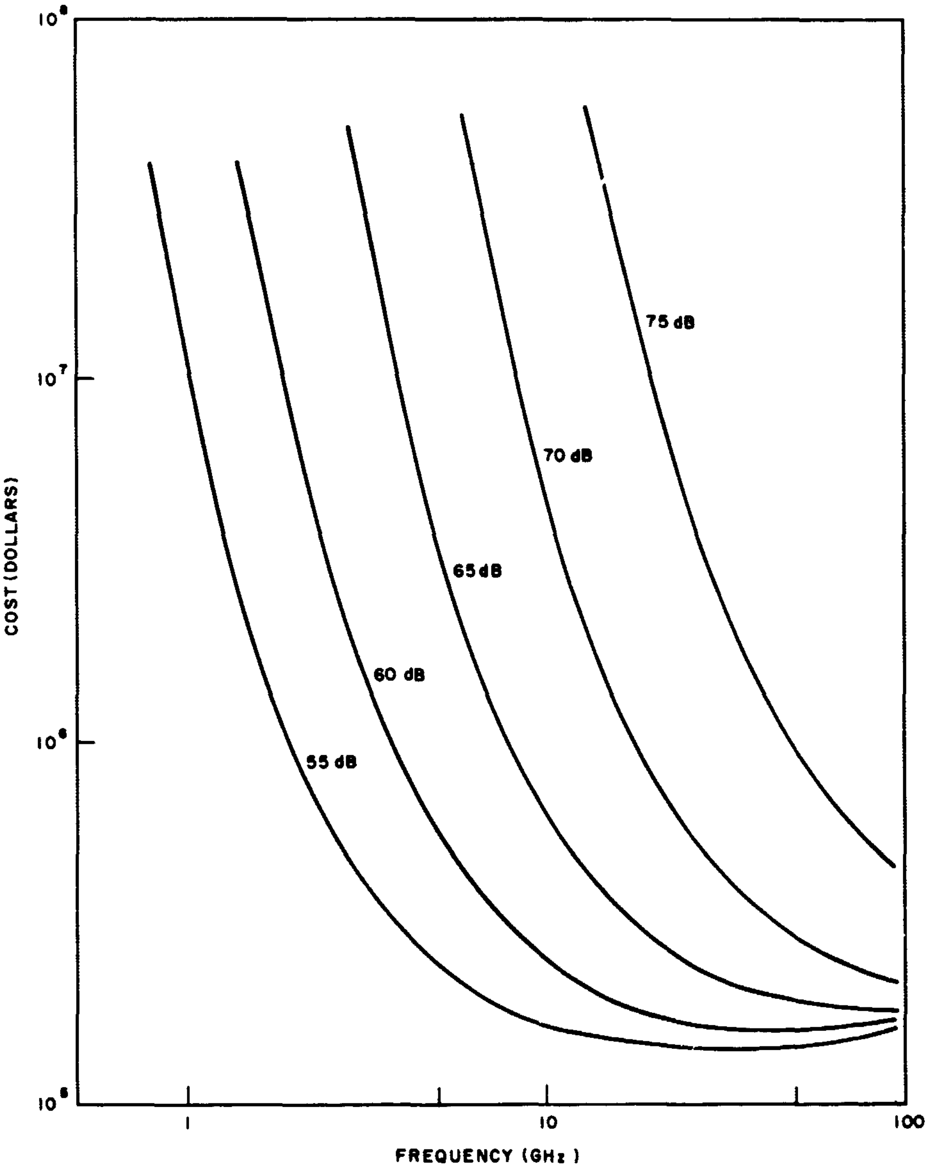


Figure 53b. Minimum cost antenna: fixed gain and frequency (enclosed by radome), cost vs. frequency

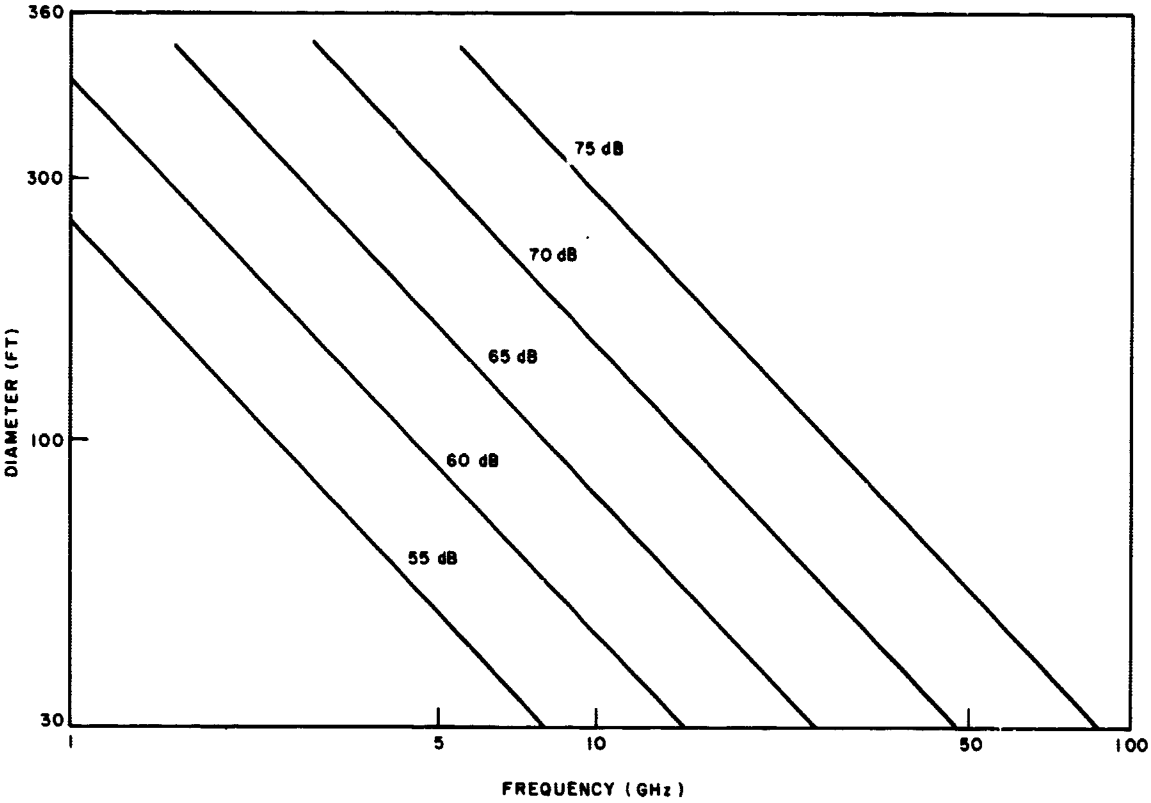


Figure 54a. Minimum cost antenna: fix·d gain and frequency (enclosed by radome), diameter vs. frequency

6. COMMUNICATION PERFORMANCE

6.1 Performance Criteria

It is not the intent of this section to provide a comprehensive review of deep-space communication theory or practice, but rather to note briefly the trade-offs that are available and to indicate the advantages and performance associated with biorthogonal modulation systems.

It will be assumed that information is transmitted from the deep-space probe to Earth by a single digital transmission link. This implies that time-division-multiplex is employed to combine the various communication channels. However, the specific details of analog-to-digital conversion and multiplexing of several channels are not within the scope of this study.

The choice of digital, as opposed to analog, modulation is prompted by several factors, most important of which are:

1. Digital tranmission allows greater flexibility than is known in analog techniques (e.g., FMFB) in achieving maximum communication rate for a given transmitter power.
2. Digital transmission facilitates computer signal processing, including correction of distorted video signals, and the use of error-control techniques.

Communication systems may generally be evaluated in terms of three criteria:

1. Power requirements
2. Bandwidth requirements
3. Fidelity.

In the case of analog communications, item 1 is usually expressed as the carrier-to-noise ratio (measured in the baseband), item 2 is the ratio of rf to base-bandwidth, and item 3 is the output signal-to-noise ratio. In the case of digital communications, the three criteria are most conveniently expressed in terms of three dimensionless quantities:

1. E/N_0
2. H/W
3. P_e

Here E is the energy per bit. If S is the average received power required to achieve an information rate H bits per second, then $E = S/H$. N_0 is the noise spectral density,* so that E/N_0 is the ratio of bit energy to noise energy per cycle of bandwidth. W is the rf bandwidth and P_e is the bit error probability.

For the case of communication in the presence of additive white Gaussian noise, a situation which prevails for space communication at microwave frequencies,† the maximum communication rate consistent with arbitrarily small error probability is given by Shannon's formula for the channel capacity:

$$H < C = W \log_2 \left(1 + \frac{S}{N_0 W}\right) \qquad (17)$$

It follows from Equation (17) that

$$E/N_0 > \frac{2^{H/W} - 1}{H/W} > \log_e 2 \qquad (18)$$

so that there is a lower bound to the required energy per bit.[90,91]

6.2 Choice of Modulation

Digital modulation consists of associating blocks of L bits of data with one of $M = 2^L$ distinct waveforms of duration T. The information rate is then given by

$$H = \frac{L}{T} = \frac{\log_2 M}{T} \qquad (19)$$

where M is termed the alphabet size; M = 2 corresponds to binary communication.

Digital modulation may, therefore, be considered as a transformation between a sequence of binary data and a voltage waveform. This is to be distinguished from coding which is a transformation from one binary data stream to another. For example, a block of L bits of data may be transformed (coded) into a block of N binary digits ($N > L$) and these may be modulated taking L' binary digits at a time, where L' may be chosen completely independent of L and N. Typically, $L' = 2$; i.e., coding is followed by binary modulation, but this is needlessly restrictive, as will be noted below. Principal emphasis will be given here to the choice of modulation, but some discussion of coding will be given in the following sections.

An M-ary modulation scheme is defined by prescribing a set of waveforms $S_i(t)$, $t = 1, \ldots, M$ on the interval (O,T) and a procedure for associating each of the M possible combinations of L bits ($M = 2^L$) with a particular S_i. The transmitted signal (apart from a possible frequency translation) is then of the form

$$\sum_n Si(n)^{(t-nT)}$$

where n indexes the blocks of L bits.

For fixed M, the optimum modulation system (in the sense of minimizing the E/N_0 required for a given error probability) is simplex modulation,[92,93] which is characterized by

*$N_0 = kT_e$ where T_e is the effective system noise temperature (degrees Kelvin) and $k = 1.386(10)^{-23}$ joules/°K is Boltzmann's constant.

†As noted in Chapter 4, the Gaussian assumption may often, but not always, be applied in the case of optical communications.

$$\int_0^T dt\, S_i(t)\, S_j(t) = \begin{cases} E_s \text{ for } i = j \\ \dfrac{-E_s}{M-1} \text{ for } i \neq j \end{cases} \quad (20)$$

Thus simplex modulation has the property that any two distinct waveforms have the same correlation coefficient. It should be noted that Equation (20) does not uniquely define the $S_i(t)$, but rather prescribes a class of systems which all yield the same performance.

A system which achieves very nearly the same power efficiency as simplex with, however, important simplifications in the generation and processing of the signals is biorthogonal modulation[93-95] which is characterized by

$$\int_0^T dt\, S_i(t)\, S_j(t) = \begin{cases} E_s \text{ for } i = j \\ -E_s \text{ for } i + j = M + 1 \\ 0 \text{ otherwise} \end{cases} \quad (21)$$

Thus, an M-ary (M even) biorthogonal modulation system employs an alphabet consisting of M/2 orthogonal waveforms together with the negatives of these waveforms. Although this might be implemented by combined pulse-position (or pulse frequency) and phase modulation, there is a particularly simple method of generating the signals as binary phase-reversal sequences.[94,95]

An example of a biorthogonal set of waveforms, for the case of M = 8, is shown in Figure 55. Each signal con-consists of four binary pulses (in general, the number of pulses is M/2). This binary waveform is used to modulate an rf carrier in such a way that a change in pulse sign corresponds to a 180-degree phase-reversal of the carrier.* In the above example, four pulses are used to transmit three bits of information. More generally, $M/2 = 2^{L-1}$ pulses are used to transmit $\log_2 M = L$ bits of information so that the ratio of pulse rate W† to information rate H is given by

$$\frac{W}{H} = \frac{M}{2 \log_2 M} = \frac{2^{L-1}}{L} \quad (22)$$

It should be noted that, although (according to the definitions given here) a modulation rather than a coding system has been described, the transmitter may be implemented in the same way that a simple shift register binary-coded system is implemented. Thus an input sequence of L bits generates an output sequence of $M/2 = 2^{L-1}$ binary digits in a Reed-Muller code generator,‡ and these output digits are used to biphase modulate an rf carrier, as illustrated in Figure 56.

"The entire waveform generating process is extremely simple and represents only a small additional complexity to any digital communication system."[95] Although receiver complexity is somewhat greater, the receiver is on the ground, and the additional complexity is by no means incommensurate with the inherent complexities of the ranging and tracking systems.

The basic elements of the receiver consist of:

1. Carrier recovery and coherent demodulation to baseband
2. Pulse synchronization
3. Word synchronization
4. Data demodulation.

Carrier recovery is obtained most simply (as mentioned above) by using less than 180-degree modulation, in which case there is a carrrier component of the transmitted waveform which may be detected in a coherent phase-lock receiver.

The achievement of pulse synchronization is similar to that of achieving bit synchronization in a binary communication system. It is made somewhat more difficult, however, by the larger pulse rates and the smaller signal-to-noise ratio with which biorthogonal systems normally operate.

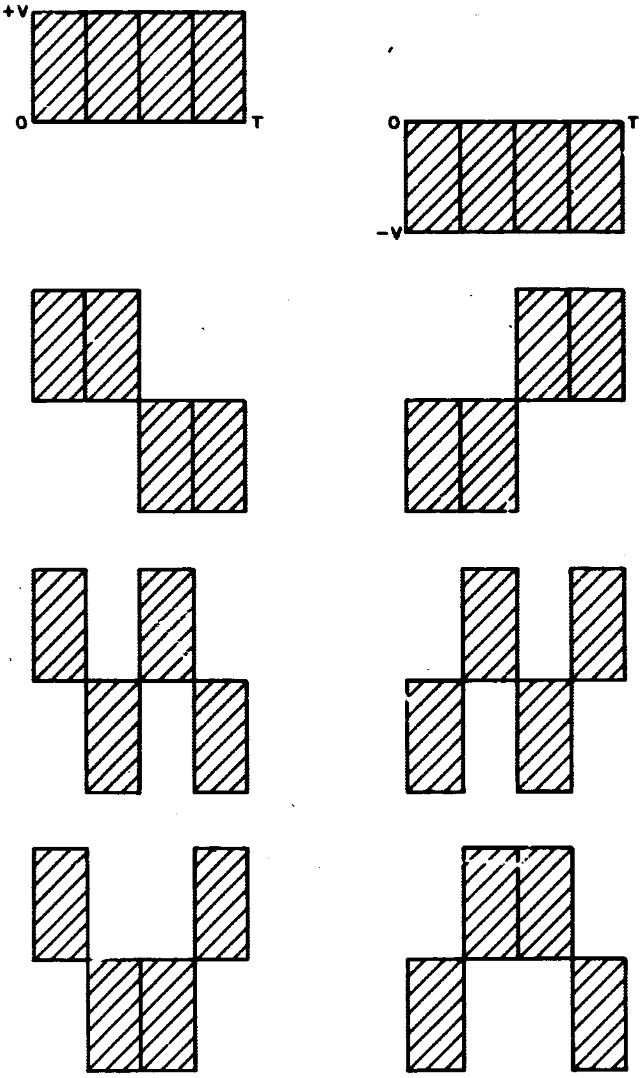
For a given information rate, the pulse rate increases almost linearly with alphabet size, as indicated by Equation (20) and shown in Figure 57. For example, the use of M = 256 (L = 8) requires a pulse rate of $16(10)^5$ p/s to achieve an information rate of 1 Mb/s which, however, is well within the capabilities of digital circuitry. Pulse synchronization needs to be obtained to within about 6 ns. Although this can in principle be obtained from the data, in practice periodic use of synchronization codes is simpler and need only take a small fraction of the available time. Such codes may also be used to achieve word synchronization.

Once pulse and word synchronization are obtained, the basic receiver operation (see Figure 56) consists of a phase detector (multiplication of the baseband signal with a synchronized square-wave clock) and an integration and dump circuit for the pulse period. Each output is fed to M accumulators which either add or subtract successive inputs according to a preset pattern determined by the Reed-Muller code. After M/2 such inputs are "added"

*This results in a DSB-SC system. Although there are techniques for recovering the carrier from the sidebands it is simpler to use slightly less than 180-degree phase shift, in which case a carrier component is present. The fraction of power devoted to the carrier may be made negligibly small in a high data rate system, since the bandwidth of the carrier tracking circuit is much less than the information bandwidth.

†Note that the pulse rate is essentially a measure of the rf bandwidth.

‡Alternate mechanizations of the Reed-Muller code generator are discussed in References 94 and 95.



F.gure 55. Octary biorthogonal waveforms

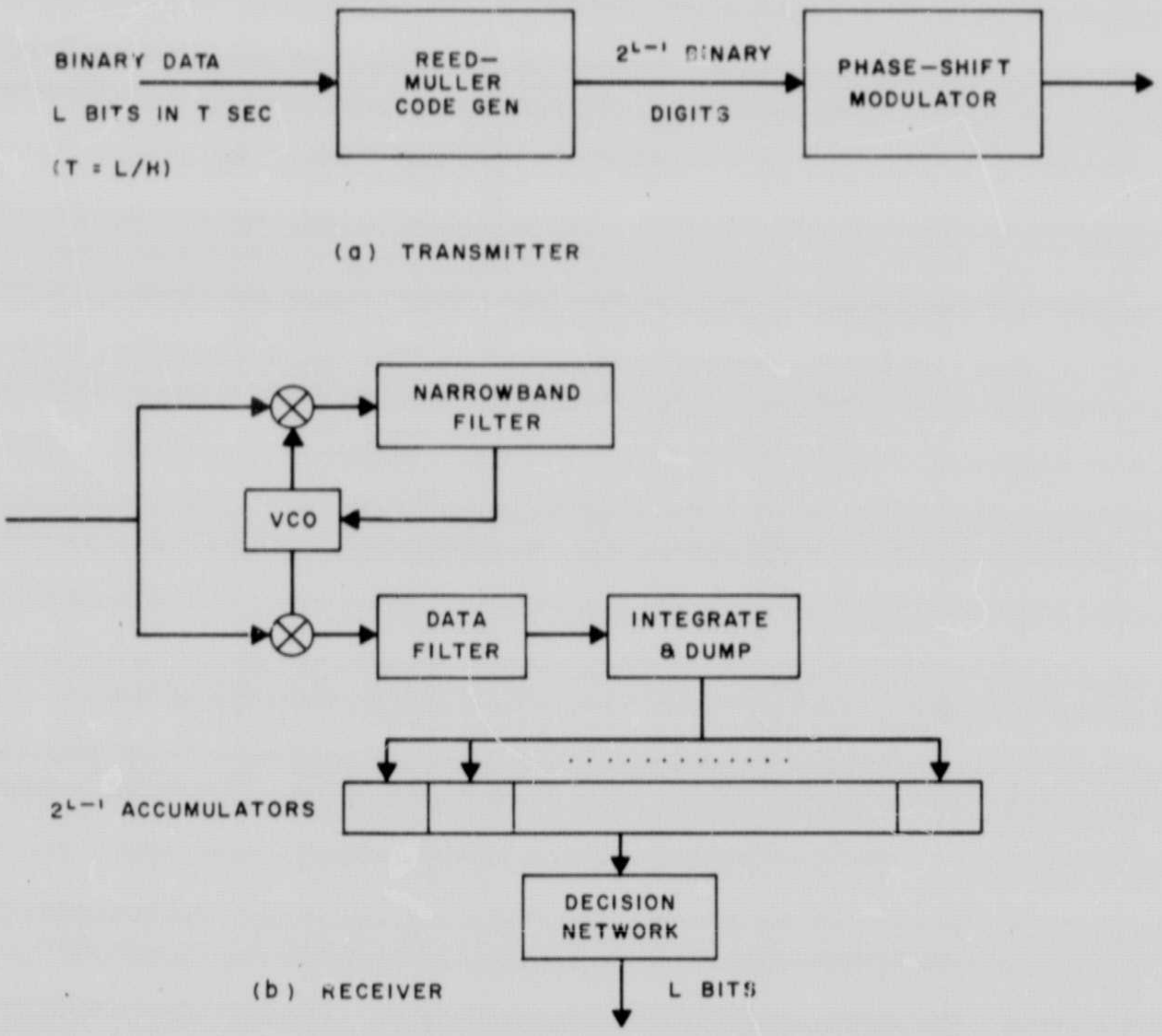


Figure 56. Biorthogonal communication system

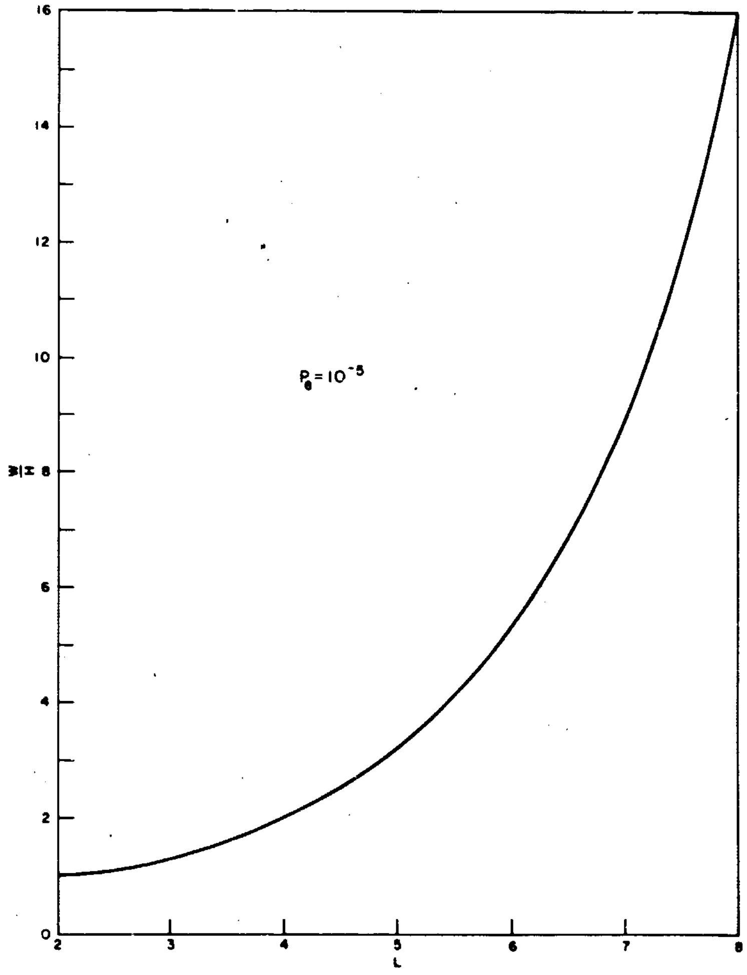


Figure 57. Cycles per bit vs. code block length, biorthogonal modulation

in each of the M accumulators, the data demodulation is obtained by determining which accumulator output is largest.* This then determines which of the M waveforms was probably transmitted, and then fixed digital circuitry converts this into a stream of L binary digits. Although the above can be implemented with delay lines and analog circuitry, in practice (particularly for large M) special-purpose digital techniques seem preferable. Since there are M accumulators, each of which has M/2 inputs, and since the largest output has to be determined, of the order of $M^3/2$ additions have to be performed.† Certainly, values of M of the order of several hundred are practical, and with parallel logic several thousands are feasible.

It is difficult to set an upper bound on M on the basis of complexity. However, there is definitely a point of diminishing returns for improvements in efficiency associated with large values of M. In Figure 58 the E/N_0 required to achieve a bit error probability of 10^{-5} (as obtained from Reference 94) is shown as a function of M. If one were to adopt the rather arbitrary requirement that a doubling of M is justified only if it achieves at least a 0.5-dB reduction in E/N_0, then Figure 58 indicates that one should stop at M = 64, for which $E/N_0 = 6.1$ db. This is 3.5 dB better than the best binary system. To achieve $E/N_0 = 5$ dB requires $M = 2^9 = 512$; to achieve $E/N_0 = 4$ dB requires $M = 2^{14} \approx 3.3(10)^5$.

The above considerations suggest a value of M between 2^6 and 2^9 with E/N_0 requirements (for $P_e = 10^{-5}$) between 5 and 6 dB.

6.3 Threshold Demodulation

An inherent difficulty with large alphabet modulation or coding techniques, when employed on high data rate channels, is that the computational-speed requirements may exceed practical limits. In the case of biorthogonal modulation, the M/2 accumulators in the detector, of course, operate in parallel. However, at the end of the word period, optimum detection requires determination of which accumulator has the largest value, which requires M/2 comparisons to be made in a time that is small compared to the bit period. Although many of these comparisons can be done in parallel, approximately $\log_2 M$ sequential operations must be performed.

It is possible to circumvent this problem, at some expense in E/N_0, by using threshold rather than optimum demodulation. For threshold demodulation, each accumulator is compared with a fixed threshold rather than with one another, and this can be done completely in parallel.

The error probability for threshold detection may be given approximately by [91]

$$P_e \approx \phi\left[-u(1-\alpha)\right] + \left(\frac{M}{2} - 1\right)\phi(-\alpha u) \qquad (23)$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} dy\, e^{-y^2/2} \qquad (24)$$

is the cumulative zero-mean unit-variance normal distribution,

$$u = \sqrt{\frac{2E}{N_0} \log_2 M} \qquad (25)$$

is the average output of the "correct" accumulator, and α u is the threshold. The first term in Equation (23) is the probability that the "correct" accumulator does not exceed the threshold, and the second term in Equation (23) is (to a good approximation) the probability that at least one of the incorrect accumulators exceeds the threshold. The fractional threshold α may be chosen to minimize P_c which essentially corresponds to making the two terms in Equation (23) equal. This results in the following expressions for the optimum threshold and the resulting enor probability

$$\alpha = \frac{1}{2} + \frac{1}{u^2} \log_e \left(\frac{M}{2} - 1\right) \qquad (26)$$

$$P_c = 2\phi\left[-\frac{u}{2} + \frac{1}{u}\log_e\left(\frac{M}{2} - 1\right)\right] \qquad (27)$$

This is the word-error probability which is then readily converted into the bit-error probability.[94]

The above results assume $M \geqslant 4$ and $P_c \ll 1$. Specifically, to achieve $P_e = 2.2(10)^{-5}$ with $M = 2^8$ (this corresponds to a bit-error probability of 10^{-5}) requires $E/N_0 = 7.6$ dB which, is 2.3 dB greater than optimum reception (see Figure 58).

The use of threshold demodulation permits an interesting option. If the threshold is chosen to be higher than the optimum value given by Equation (26), then the likelihood of an incorrect accumulator exceeding the threshold is diminished, at the expense of an increased probability that the correct accumulator will be below threshold. If, however, the latter occurrence is considered to be a "deletion" rather than an error, then the allowance of a deletion probability in excess of the desired error probability can result in considerably reduced E/N_0. This is illustrated in Figure 59, where deletion probability is shown as a function of E/N_0 for fixed-bit error probability of 10^{-5} for $M = 2^8$ and $M = 2^{12}$. It is seen that if a 1-percent deletion probability is allowed, then the E/N_0 requirements are 5.4 and 4.3 dB, respectively. If a 10-percent deletion

*Alternatively, because of the biorthogonal property, only M/2 accumulators need be employed, and the decision is then made on the basis of the sign of the accumulator whose output has the largest absolute value.

†Since the number of operations increases as a power of M, it is exponentially growing in L. This is characteristic of "block" coding and modulation systems.

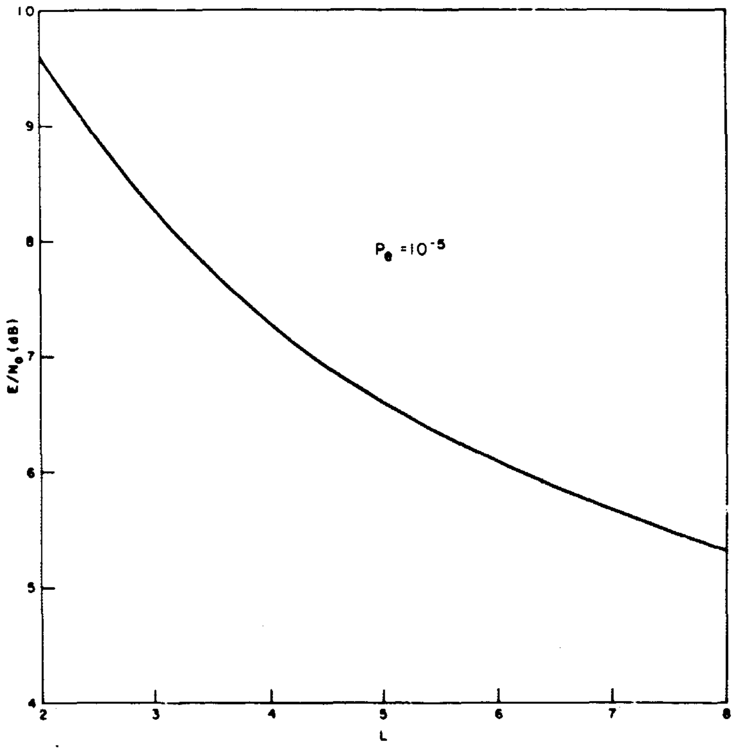


Figure 58. Signal-to-noise ratio vs. code block length (2^{L-1} binary digits) for biorthogonal modulation system

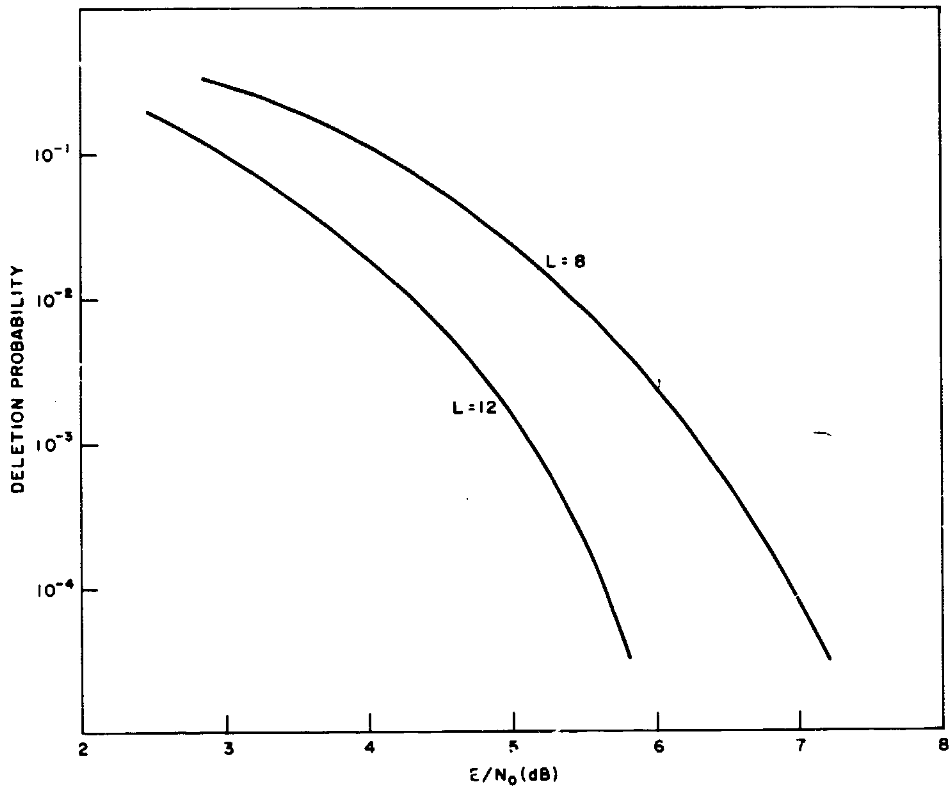


Figure 59. Bioi thogonal modulation with threshold detection

probability is allowed, the corresponding numbers are 4.1 and 3 dB. Thus, by allowing a modest deletion rate, a significant reduction in E/N_0 may be achieved.

It is possible, of course, to apply the optimum detection procedure to that small fraction of the data for which a deletion occurs.* If the deletion rate is small, the computational requirements for this operation can be kept modest. It should be noted, of course, that the combination of threshold and optimum detection cannot yield better performance than optimum detection applied to the entire data. However, the former procedure can make practical much larger alphabet sizes than are practical with optimum detection, and in that sense can afford significant performance improvements.

6.4 Convolutional Coding

The biorthogonal modulation system discussed above may also be considered as a special case of a block code. Both the optimum and threshold demodulation described in Section 6.3 differ from conventional decoding in that hard decisions are not made on the basis of individual received pulses; rather, the analog voltages in each sampling interval are accumulated. If the receiver is to be implemented digitally, then these voltages must be quantized. Although the above results apply strictly to the case of infinite quantization, in practice 8-level quantization generally comes to within a small fraction of a decibel of the infinite quantization result.[96] On the other hand, 2-level quantization (hard decisions on the basis of each pulse) generally entails a 2-dB penalty.

In a block code, successive blocks of L bits of data are transformed into successive blocks of N binary digits, and there is no interrelation between blocks. Convolutional codes, on the other hand, achieve a sliding dependence in which each bit of input data affects Lv digits of the output sequence where L is the constraint length of the code and v is the redundancy. A convolutional coder consists simply of an L-bit shift register with v modulo two adders as indicated in Figure 60. Each time a new bit enters the shift register, the v adders are sampled to produce v output digits. Typically v = 2 or 3 corresponds to an output binary digit rate 2 or 3 times the input rate.

It should be noted that shift register and adders might also be employed to generate a block code. If L bits are to be encoded into N bits, then N adders are required. Since N typically is a large number, whereas v is typically a small number, the convolutional coder offers important implementation advantages.

Since convolutional coding does not divide the input and output sequences into independent blocks, optimum decoding requires observation of the entire message. Sequential decoding is a probabilistic procedure for avoiding the exponential growth feature of optimum decoding.

In a convolutional encoder, each time a new input bit is added, one of two branches on a code tree is followed. For L input bits there are 2^L branches. The purpose of a sequential decoder is to proceed through this tree examining substantially fewer than the totality of branches. Unlike block decoding, the number of branches examined, and hence the number of computations which must be performed, is a random variable. In practice, the number of branches which can be examined before proceeding one node is limited by the computational speed of the decoder. If this number is exceeded, it is impossible to decode and this may be considered (as in the previous section) as causing a deletion. Recent results[97,98] indicate that to achieve deletion probabilities of the order of 1 percent may require the ability to examine several hundred branches.† It is questionable whether this is feasible at megabit rates.

Several recent papers[97-99] have described, evaluated, and advocated the use of sequential decoding for space application, but specific implementations[98,99] are limited to data rates in the kilobit range. Jacobs[97] indicates that "the application of sequential decoding to space communications is very attractive and promises operation within 3 to 4 dB of the ultimate limit set by channel capacity.... In practice, the difficulty in reducing P_0 (deletion probability) to a truly small value is the ultimate limit on the performance of systems with sequential decoders."

Although sequential decoding appears to yield perhaps 2 dB smaller E/N_0 than biorthogonal (or other block codes), it appears to have two disadvantages relative to the biorthogonal system:

1. The variable computational requirements make it difficult to engineer a high data rate system.
2. Sequential decoding requires that the bulk of the computation be done sequentially, whereas biorthogonal (or other block-coded) systems allow considerable parallel operations.

The intention is not to advocate a specific modulation or coding system here but rather to indicate that there are systems which achieve improvements of the order of 4 to 6 dB relative to uncoded binary modulation systems. Detailed comparisons of different modulation and coding techniques are outside the scope of this study. However, the limits will probably be imposed by the computational requirements on the receiver and, if moderate deletion rates are allowed (in either block or convolutional codes), the computational problem may be greatly reduced.

*The use of a feedback channel with request for retransmission may not be practical because of the long transmission delay and consequent storage requirements imposed on the transmitter.

†Although systematic procedures have been developed for both decoding algorithms[96] and for obtaining good codes,[98,99] unfortunately the determination of performance is still largely on the basis of simulation rather than analysis, and it is difficult (at least for those not well versed in the field) to draw general conclusions.

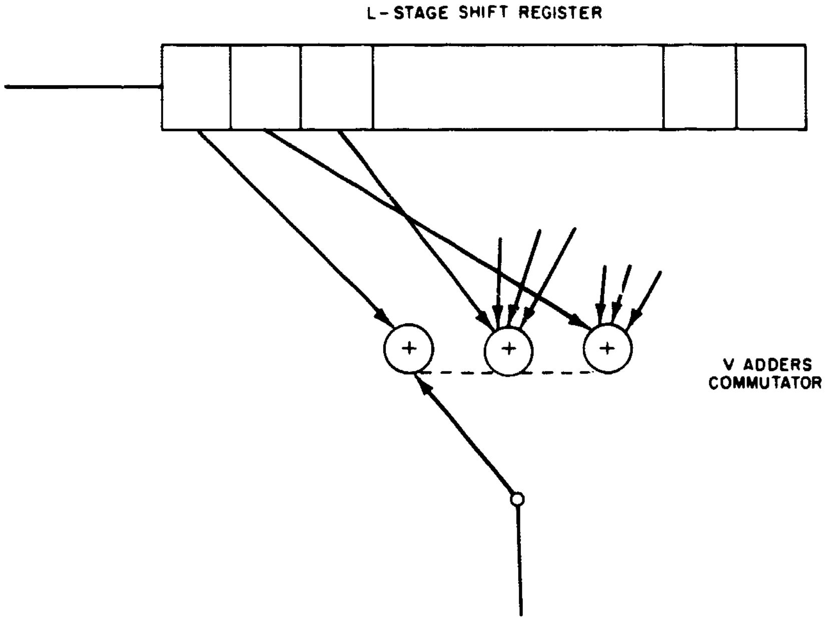


Figure 60. Convolution coder

REFERENCES

1. R.W. Hartop, Advanced Development of Microwave Antenna Subsystems, SPS 37 – 40, Vol. III, Jet Propulsion Laboratory, Pasadena, California (July 31, 1966).

2. P.D. Potter, "Application of Spherical Wave Theory to Cassegrainian-Fed Paraboloids," IEEE Transactions on Antennas and Propagation, Vol. AP-15 (November, 1967), pp 727-736.

3. R.W. Hartop, Advanced Development of Microwave Antenna Subsystems: 100-KW SCM Cone Tests, SPS 37 – 41, Vol. III, Jet Propulsion Laboratory, Pasadena, California (September 30, 1966).

4. J.F. Hull, "Microwave Tubes of the Mid-Sixties," IEEE Intl. Conv. Rec., Part 5 (March, 1965), pp 67-78.

5. J.M. Osepchuk, "Toward a Renaissance in Microwave Tubes," Microwave Journal (September, 1967), pp 18 ff.

6. B.C. Deloach, "Recent Advance in Solid State Microwave Generators," Advances In Microwave, Vol. 2, Leo Young, Editor, Academic Press (1967) pp 43 – 88.

7. N.E. Feldman, "Communication Satellite Output Devices-Part II," Microwave Journal (December, 1965), pp 87-97.

8. Private communication with Wesley Teich, Raytheon Company, Microwave and Power Tube Division, Waltham, Massachusetts.

9. J.T. Mendel, "High-Powered Traveling-Wave Tubes for Space Transmitters," Communication Satellite Systems Technology, R.B. Marsten, Editor, Academic Press (May, 1966), pp 529-548.

10. M. J. Schindler, "Advances in Traveling-Wave Tubes for Spacecraft Communications Systems," Communication Satellite Systems Technology, R.B. Marsten, Editor, Academic Press (May, 1966), pp 423–432.

11. Parametric Analysis of Microwave and Laser Systems for Communication and Tracking, Phase I Final Report, Hughes Aircraft Company, Contract No. NAS 5-9637 (February, 1966).

12. L.A. Roberts and M.V. Purnell, Development of a 100 Watt S-Band Traveling-Wave Tube, Watkins-Johnson Co., JPL Contract No. 951299 (April 21, 1967).

13. E.L. Lien and A. Mizuhara, ESFK 20-100 Watt S-Band Amplifier, Eimac, JL Contract No. 951105 (July, 1967).

14. J.T. Mendel, "Progress in TWT Development," Intl. Electronics (October, 1966), pp 38-42.

15. M.J. Schindler, "Can Traveling Wave Tubes Be Scaled?" Microwave Journal Vol. 9 (September, 1966), pp 43-47.

16. E.L. Lien, A. Mizuhara, and D.I. Boilard, Electrostatically-Focused Extended-Interaction S-Band Klystron Amplifier, presented at Intl. Electron Devices Meeting, Washington, D.C. (October 26, 1966).

17. A.J. Prommer, Electrostatically Focused Klystron Amplifiers for Deep Space Communications, presented at the Intl. Space Electronics Symposium, Miami Beach, Florida (November 2-4, 1965).

18. D.H. Preist, "The Future of Extended Interaction Klystrons" Proc. 5th Int. Congress on Microwave Tubes (1963), p 289.

19. D.H. Priest and W.J. Leidigh, "A Two-Cavity Extended Interaction Klystron Yielding 65 Percent Efficiency," IEEE Transactions on Electron Devices, Vol. ED-11 (August, 1964), pp 369-373.

20. D.H. Priest and W.J. Leidigh, "Experiments with High Power CW Klystrons with Extended Interaction Catchers," IEEE Transactions on Electron Devices, Vol. ED-10 (May, 1963), pp 201-211.

21. M. Chodorow and T. Wessel-Berg, "A High Efficiency Klystron with Distributed Interaction," IEEE Transactions on Electron Devices, Vol. ED-8 (January, 1961), pp 44-55.

22. J.A. Copeland, "CW Operation of LSA Oscillator Diodes – 44 to 88 GHz," BSTJ, Vol. 46 (January, 1967), pp 284-287.

23. N.E. Feldman, "Communication Satellite Output Devices – Part I," Microwave Journal, Vol. 8 (November, 1965), pp 69-72.

24. M.R. Barber, Solid-State Devices for Microwave Power Generation, paper presented at 1967 International Solid-State Circuits Conference, Philadelphia, Pennsylvania (February 17, 1967).

25. D.M. Snider, "High Power High Efficiency Transistor RF Amplifiers," 1967 NEREM Record, pp 20-21.

26. R.D. Brooks and J.W. Gewartowski, "A Unilateral 6 GHz 2-1/2-Watt Varactor Quadrupler," 1967 International Solid-State Circuits Conference, Digest of Technical Papers.

27. Satellite Control Satellite Planning Analysis, proposal to USAF Headquarters Space and Missile Systems Organization, August 14, 1967.

28. R.M. Ryder, Note on Prospects for Future Development of Solid-State Power Sources, private correspondence to W.L. Mraz, Bell Telephone Laboratories, May 17, 1967.

29. J.A. Copeland and R.R. Spiwak, LSA Operation of Bulk n-GaAs Diodes, paper presented at 1967 International Solid-State Circuits Conference, Philadelphia Pennsylvania, February 15, 1967.

30. Private communication with J.R. Neverez, Bell Telephone Laboratories.

31. O.G. Sauseng and M.N. Ernstoff, Advancements in TWT Efficiency with Combined Voltage Jump, Velocity Tapers and Enhanced Beam Bunching, presented at International Electron Devices Meeting, Washington, D.C., October, 1966.

32. Applied Research on Efficiency Improvement in O-Type Traveling Wave Tubes, Final Report, Bell Telephone Laboratories, Contract No. AF33 (615)-1951 (April, 1967).

33. B.A. Highstrete, "Miniaturized Metal-Ceramic Traveling Wave Tubes for Space Vehicles," Microwave Journal (January, 1964), pp 49-57.

34. R.A. Brenan and N.A. Greco, Ultra High Reliability Traveling-Wave Tubes, presented at the Fifth AIAA-SAE/ASME Reliability and Maintainability Conference, New York, New York (July 18-20, 1966).

35. M.G. Bodmer et al, "The Satellite Traveling Wave Tube," BSTJ (July, 1963), pp 1703-1748.

36. Study of High Power Transmitting Tubes Above X-Band, Final Report, Bell Telephone Laboratories USAECOM Contract No. DA-28-043AMC-00234 (E) (July, 1965).

37. T.B. Brown and A.L. Rousseau, 10 kW Traveling Wave Tubes for CW Communication at C-Band, X-Band, and KU-Band, presented at International Electronic Devices Meeting (October 29, 1964).

38. N.W. Snyder, "Power Systems," Astronautics (November, 1962), pp 110-114.

39. G.C. Szego, "Space Power Systems State of the Art," Journal of Spacecraft and Rockets, Vol. 2 (September-October, 1965), pp 641-659.

40. F.L. Raposa, "Non-Dissipative DC-to-DC Regulator Convertors," EASTCON 67 Technical Convention Record, pp 410-417.

41. J.W. Bates, Power Conditioning System Design, Paper 20.1, presented at WESCON (1965).

42. J.P. Powell et al, "Power Conditioning Development for Ion Engines," Space Power Systems Engineering, G.C. Szego, J.E. Taylor, Editors (Academic Press, 1966), pp 1261-1271.

43. D.L. Southern, "Power Systems Comparison For Manned Space Station Applications," Space Power Systems Engineering, G.C. Szego, J.E. Taylor, Editors, Academic Press (1966), pp 335-359.

44. Private communication with J.F. Heney, Hughes Research Labs, Malibu, California.

45. Private communication with L.E. Bernier, Litton Industries, San Carlos, California.

46. A. Lubarsky et al, "S- and X-Band Traveling-Wave Tube Amplifiers for Satellite Applications," Wescon Technical Papers, Vol. 8, 1964 (presented August 25-28, 1964).

47. Private communication with Dr. J. Lucas, Jet Propulsion Laboratory, and H.J. Scagnelli, Bell Telephone Laboratories, July 28, 1967.

48. W.H. Woodard, NASA, statement before the Committee on Aeronautical and Space Sciences, United States Senate.

49. A. Basiulis and M.C. Starr, Improved Reliability of TWT's Through the Use of a New Lightweight Heat Removal Device, paper presented at International Electron Devices Meeting (October 19, 1967).

50. H.E. King, "Rectangular Waveguide Theoretical CW Average Power Rating," IRE Trans. on Microwave Theory and Techniques, Vol. MTT-9 (July, 1961), pp 349-357.

51. J. Ruze, "Antenna Tolerance Theory-A Review," Proceedings of the IEEE, Vol. 54, No. 4 (April, 1966), pp 633-640.

52. J.W. Mar and F.Y.M. Wan, "The Influence of Shell Behavior on The Design of Large Antennas," Proc. XVI Internat'l Astronaut. Cong., Athens, Greece (1965), pp 183-212.

53. S. von Hoerner, "Design of Large Steerable Antenna," The Astronomical Journal, Vol. 72, No. 1 (February, 1967), pp 35-47.

54. Study of Conceptual Deep Space Monitor Communications Systems Using a Single Earth Satellite, Vols. I, II, III; SGC 920FR-1; Report of Mission Analysis Division, NASA, Moffett Field, California (NASA Contract 2-3197), prepared by Space General Corporation, El Monte, California (September, 1966).

55. Feasibility Study of Large Space Erectable Antennas, Vol. I, II; GDC DCL 67-002; Report to Navigation and Communication Division, NASA Headquarters, Washington, D.C. (NASA Contract NAS W 1438), prepared by Convair Division of General Dynamics, San Diego, California (April, 1967).

56. An Advanced Study of an Application Technology Satellite (ATS-4) Mission, Vol. I, Bk. 2, Final Report, General Electric Company, Missile and Space Division, NASA-CR-81767 (November, 1966).

57. ATS-4 Study Program, Vol. 3, Final Report, Fairchild-Hiller Corporation, Space Systems Division, NASA-CR-81603 (December, 1966).

58. Advanced Study of an Applications Technology Satellite (ATS-4) Mission, Final Report, Lockheed Missiles and Space Company, NASA-CR-81765 (November, 1966).

59. I.D. Smith, "Role of Millimeter Waves in Deep Space Telecommunication Systems," 1966 National Telemetering Conference Proceedings, Boston, Massachusetts (May, 1966), pp 307-313.

60. D.C. Hogg, "Communication Through the Atmosphere at Millimeter Wavelengths," Science, Vol. 159 (January 5, 1968), pp 39-46.

61. R.A. Semplak, "Gauge for Continuously Measuring Rate of Rainfall," Rev. of Sci. Instru., Vol. 37 (November, 1966), pp 1554-1558.

62. E.A. Mueller and A.L. Sims, Illinois State Water Survey, under Contracts SC-75055 and SC-87280 with the U.S. Army Signal Corps (to be published in Journal of Meterology, AMS).

63. J.W. Ryde and D. Ryde, Attenuation of Centimetre Waves by Rain, Hail and Clouds, General Electric Co., Wemblay, England, Report 8516 (August, 1944).

64. J.W. Ryde and D. Ryde, Attennation of Centimetre and Millimetre Waves by Rain, Hail, Fogs and Clouds, General Electric Co., Wemblay, England, Report 8670 (May, 1945).

65. J.O. Laws and D.A. Parsons, "The Relationship of Raindrop-Size to Intensity," Trans. Am. Geophys. Union, Vol. 24 (1943), pp 452-460.

66. D.C. Hogg, "Effective Antenna Temperatures due to Oxygen and Water Vapor in the Atmosphere," J. of Appl. Phys., Vol. 30 (September, 1959), pp 1417-1419.

67. D.C. Hogg, "Ground-Station Antennas for Space Communication," Advances in Microwaves, Vol. 3, Academic Press (to be published).

68. D.E. Kerr, Propagation of Short Radio Waves (New York, McGraw-Hill Book Co., 1951).

69. K.L.S. Gunn and T.W.R. East, "The Microwave Properties of Precipitation Particles," J. Roy. Meteorol. Soc., Vol. 80 (1954), pp 522-545.

70. D.C. Hogg, "Path Diversity in Propagation of Millimeter Waves Through Rain," IEEE Transactions, Vol. AP-15 (May, 1967), pp 410-415.

71. R.G. Medhurst, "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Measurement," IEEE Transactions, Vol. AP-13 (July, 1965), pp 550-564.

72. B.C. Blevis et al, "Measurements of Rainfall Attenuation at 8 and 15 GHz," IEEE Transactions, Vol. AP-15 (May, 1967), pp 394-403.

73. D.C. Hogg and R.A. Semplak, "The Effect of Rain and Water Vapor on Sky Noise at Centimeter Wavelengths," BSTJ, Vol. 40 (September, 1961), pp 1331-1348.

74. K.N. Wulfsburg, "Sky Noise Measurements at Millimeter Wavelengths," Proc. IEEE, Vol. 52 (March, 1964), pp 321-322.

75. K.N. Wulfsburg, "Atmospheric Attenuation at Millimeter Wavelengths," Radio Science, Vol. 2 (March, 1967), pp 319-324.

76. D. Gibble, "Sky Noise Measurements at 5.35 GHz During Rainfall," to be published in BSTJ.

77. J.S. Marshall, C.D. Haltz, and M. Weiss, McGill University Scientific Report MW-48 (ANTC Report No. 109) (May, 1965).

78. E.K. Smith and S. Weintraub, "The Constants in the Equation for Atmospheric Refractive Index at Radio Frequencies," Proc. IRE, Vol. 41 (August, 1953), pp 1035-1037.

79. B.R. Bean and B.A. Cahoon, "The Use of Surface Weather Observations to Predict the Total Atmospheric Bending of Radio Rays at Small Elevation Angles," Proc. IRE (Correspondence), Vol. 45 (November, 1957), pp 1545-1546.

80. B.R. Bean, "Tropospheric Refraction," Advances in Radio Research, Vol. 1 (Academic Press, 1964).

81. P.E. Schmid, Atmospheric Tracking Errors at S- and C-Band Frequencies, NASA Technical Note TND-3470 (August, 1966).

82. J.H.W. Unger, "Random Tropospheric Angle Errors in Microwave Observations of the Early Bird Satellite," BSTJ, Vol. 51 (November, 1966), pp 1439-1474.

83. J.T. Kennedy and J.W. Rosson, "Use of Solar-Radio Emission for the Measurement of Radar Angle Errors," BSTJ, Vol. 41 (November, 1962), pp 1799-1812.

84. J. Ruze, "The Effect of Aperture Errors on the Antenna Radiation Pattern," Suppl. al Nuevo Cimento, Vol. 9, No. 3 (1952), pp 364-380.

85. Private communication with B.R. Stack (July 26, 1967).

86. B.R. Stack, An Approximate Expression for the Cost of Large Parabolic Antennas, Stanford Research Institute, Menlo Park, California (in preparation).

87. P.D. Potter, W.D. Merrick, and A.C. Ludwig, "Big Antenna Systems for Deep-Space Communications," Astronautics and Aeronautics, Vol. 4, No. 10 (October, 1966), pp 84-95.

88. A Large Radio-Radar Telescope – CAMROC Design Concepts, Vol. I and II, The Cambridge Radio Observatory Committee, Cambridge, Massachusetts (January 15, 1967).

89. Design of Earth Terminals for Satellite Communications, Vol. 2, Section II, "Antennas," report to Communications Satellite Corporation, prepared by Bell Telephone Laboratories, Whippany, N.J. (April 1, 1965).

90. R. W. Sanders, "Communication Efficiency Comparison of Sevaral Communication Systems," Proc. IRE, Vol. 48 (May, 1960), pp 575-588.

91. I. Jacobs, "Theoretical and Practical Limitations of Weak-Signal Processing Techniques," Space Research II (North Holland Publishing Co., 1961), pp 413-425.

92. H.J. Landau and D. Slepian, "On the Optimality of the Regular Simplex Code," BSTJ, Vol. 45 (October, 1966), pp 1247-1272.

93. I. Jacobs, "Comparison of M-ary Modulation Systems," BSTJ, Vol. 46 (May-June, 1967), pp 843-864.

94. A.J. Viterbi, "On Coded Phase-Coherent Communications," IRE Trans. Space Elec. Tel., SET-7, No. 1 (March 1961), pp 3-14; See also Chapter 7 of Digital Communications with Space Applications, by S.W. Golomb et al (Englewood Cliffs, Prentice Hall, 1964).

95. R.W. Sanders, "The Digilock Orthogonal Modulation System," Advances in Communication Systems: Theory and Application, Vol. 1, Chapter 3, A.V. Balakrishnan, editor (Academic Press, 1965).

96. J.M. Wozencraft and I.M. Jacobs, Principles of Communication Engineering, Chapter 6 (New York, John Wiley & Sons, Inc., 1965).

97. I.M. Jacobs, "Sequential Decoding for Efficient Communication from Deep Space," IEEE Transactions on Communication Technology, Vol. COM-15, No. 4 (August, 1967), pp 492-501.

98. I.L. Lebow and P.G. McHugh, "A Sequential Deoding Technique and Its Realization in the Lincoln Experiment," IEEE Transactions on Communication Technology, Vol. COM-15, No. 4 (August, 1967), pp 477-491.

99. D.R. Lumb and L.B. Hofman, An Efficient Coding System for Deep Space Probes with Specific Application to Pioneer Missions, NASA TN D-4105 (August, 1967).

CHAPTER 2. MILLIMETER SYSTEMS

The characteristics of millimeter wave, deep space communication systems are considered in this chapter. Space vehicle transmitting devices, propagation effects, and ground station receiver capabilities are considered. Antennas for both ground station and space vehicle applications at millimeter frequencies are discussed, along with the microwave antennas, in Chapter 1.

Information in this chapter is concentrated at frequencies near 35 GHz and 94 GHz because of the transmission windows (relatively low, clear-sky attenuation) that exist through the atmosphere. The high attenuation at other millimeter frequencies justifies their elimination for the study of communications from an Earth-based ground station to a deep space probe. Previous studies have indicated a relatively high atmospheric attenuation at 94 GHz under conditions of moderate rain and cloud conditions. Therefore, studies in the millimeter wave region were primarily focused at frequencies in the 35 GHz atmospheric window.

Space vehicle transmitting devices are considered first. Future device capabilities and limitations in terms of power, efficiency, lifetime, and weight at 35 GHz are discussed.

Next the atmospheric effects on millimeter wave propagation are discussed under conditions of clear sky, rain, snow, and ice. A comparison of the relative effects at 35 and 94 GHz is made. Refraction effects are also discussed. It is shown that planewave phasefront distortion is severe, at low elevation angles in particular, at 35 and 94 GHz.

Receiver noise temperature capabilities are considered for frequencies from 1 to 100 GHz. The data are oriented toward ground station use; however, some of the devices discussed may be used for spacecraft receivers. Present and projected ground station system noise temperatures (including the antenna, clear sky, and receiver noise) are also obtained and plotted.

The final section considers the effects of atmospheric turbulence, attenuation, and sky noise on communication performance at millimeter frequencies. Limits on antenna diameter, caused by atmospheric phasefront distortion, are obtained at 35 and 94 GHz. Rain and cloud attenuation are compared at various frequencies based on previously measured attenuation distributions. Since it appears that operation of a millimeter system under all weather conditions is not feasible, the alternative of specifying system margins for required operating conditions (zenith angle and allowable percent outage time) is proposed as a logical method for system comparisons.

1. SPACE TRANSMITTERS

Space transmitters at millimeter frequencies, as considered in this section, include oscillators and amplifiers at 35 and 94 GHz. Much of the groundwork for this section may be found in Chapter 1, Section 2. Practical and theoretical device limitations are considered and the performance characteristics of a millimeter-wave space transmitter for the 1980 period are specified.

Recent surveys have adequately covered the new techniques and devices for transmitters at millimeter frequencies.[1,2] Others have previously considered space transmitters to some degree. Advantages and disadvantages of various devices have been considered[3] and useful data have been listed.[4]

1.1 Device Considerations

Linear beam devices dominate the millimeter frequency range. Many low-power devices with poor efficiencies fall in this class and will not be considered.

The devices which offer the most promise at millimeter frequencies are quite similar to the extended interaction devices which dominate the higher microwave region – the extended interaction klystron and the coupled cavity TWT's.[5,6] Because these structures are extremely small at millimeter-wave frequencies, fabrication tolerances and heat provide ultimate power restrictions. However, a great amount of design effort has gone into the development of machining techniques, tolerance control, and ways of obtaining precise beam optics.[5,7,8] As a result, the power obtained by use of these "conventional" techniques appears to be more than adequate to satisfy the space communication problems considered in this report.

1.1.1 Power

Impressive powers have been reported for coupled cavity traveling wave tubes at frequencies from 35 to 94 GHz.[1,5-8] Many of these devices have been built in the 50 to 55 GHz range. To scale these results to 35 GHz a reasonable scaling law must be assumed for each device. The actual dimensional scaling of a TWT follows a complex set of conditions, but it has been shown[9] that a TWT can be scaled according to a Pf^2 scaling law. In particular, scaling from a higher to a lower frequency is feasible.

In Figure 61 several coupled cavity TWT's are plotted on a power-frequency scale. The point plotted as a square represents a recent estimate[10] suggesting that a 20 watt TWT, at 55 GHz, suitable for long-life space operation could be built (from existing designs) in one year. This point was scaled by Pf^2 to 35 GHz (the triangular point) and indicates a present capability of approximately 50 watts. If the other 55 GHz tubes (813H and 819H) are scaled, very high powers (as much as 10 kW) would be predicted at 35 GHz. These same conclusions can be reached by assuming nearly any reasonable scaling law. At present these very high power tubes are useful only in the laboratory with elaborate cooling techniques. Much is to be learned about cooling techniques in a space environment. It is likely that the capabilities of the cooling system will necessitate limiting the power output of the tubes.

In addition to the necessity of cooling the tube at 35 GHz, the standard rectangular waveguide at the output of the tube may also have to be cooled because of attenuation in the waveguide walls. Without accurately specifying the parameters in the heat transfer problem, it is impossible to determine exactly when cooling may be required; however, estimates can be made in the manner previously described (Chapter 1, Section 2.4). As an example of the calculation for 35 GHz waveguide, the CW power is limited to 470 W (assuming the same conditions as applied in Chapter 1). By adding larger radiating fins to the waveguide, a higher power limit (possibly 1 kW) is likely. Depending on the spacecraft design, some active means to conduct heat to an exterior radiating surface may be required. Thus, it appears that a reasonable power limit for the waveguide at 35 GHz is about 1 kW.

Such a goal for the tube should also be feasible and should require only a moderate cooling system capacity. Although 1 kW may be well below the power capability of the tube, this limit will simplify the reliable lifetime problem.

1.1.2 Efficiency

The efficiency of the coupled cavity TWT's is impressive compared with other millimeter wave devices. In fact, reported efficiencies as high as 35 percent are comparable to the microwave devices.

Efficiencies of 50 to 60 percent have been demonstrated[11,12] for X band tubes which are in the same basic coupled cavity family as the millimeter tubes in Figure 61. Since the millimeter tubes are so much smaller than the X band tubes, the techniques used to obtain these high efficiencies will be difficult to apply at the millimeter wave frequencies. Thus, an overall frequency of 40 percent within 10 years is a reasonable goal and a likely limit. This will be particularly true if the focusing structure is a solenoid and thus requires additional power (see the discussion of permanent magnets vs. solenoids in Chapter 1, Section 3.2).

1.1.3 Lifetime

Reported lifetimes for millimeter wave tubes usually correspond to the theoretical operating lifetime of the cathode. Thus the lifetime is distinct from the reliability of the tube (as discussed in Chapter 1, Section 2.3.4).

For the family of tubes in Figure 61, there is a large amount of lifetime information. A comprehensive study[5] produced designs for four tubes – two each at 15.5 and 31.65 GHz – with power levels of 10 kW and 25 kW at each frequency. These tubes were designed for a 10,000 hour lifetime. Tube construction is very rigid and should be inherently reliable. If precautions are taken in the manufacturing processes to avoid impurities, a 15,000 hour reliable lifetime for a 1 kW tube at 35 GHz should be possible.

There is an important qualification to the projected reliable lifetime. Millimeter tube development has been very slow. This trend is not sufficient to obtain a long reliable lifetime by 1980. To do so would require a specific program of several years' duration.

1.1.4 Weight

The largest contribution to the weight of the coupled cavity tubes of TWT's is made by the structure used for focusing. The required magnetic fields cannot be supplied by periodic magnets. The weight of the solenoid or permanent magnet required for a 1 kW 35 GHz tube is approximately 30 to 40 lb at present. This is already higher than that predicted by the weight-power relationship for the microwave tubes in Chapter 1, Section 2.3.3 (in Figure 20, W = 10 lb for P = 1 kW). Recent efforts to

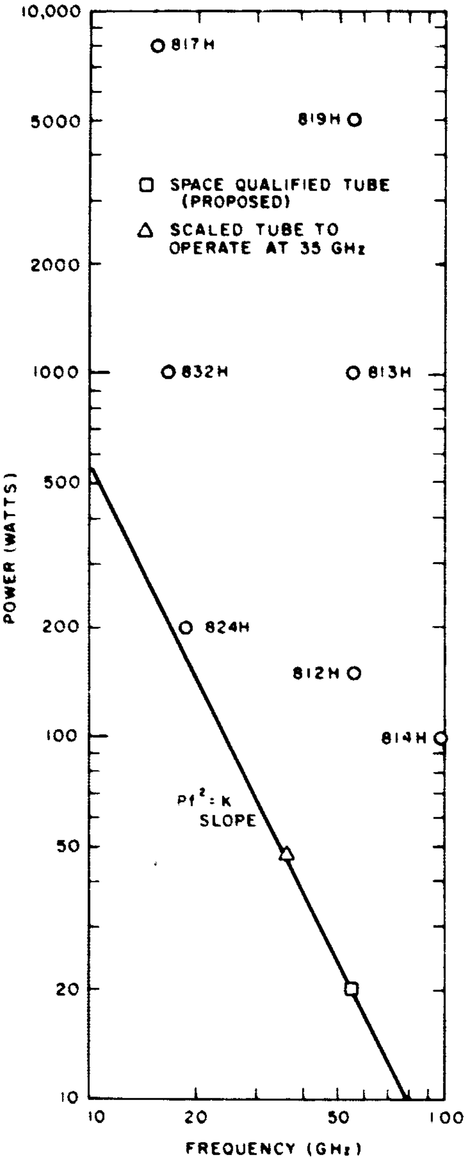


Figure 61. High power millimeter wave tubes

reduce solenoid weight by approximately one-half have been successful but the cost has been added complexity in the cooling system required for the solenoid.

There is no space-qualified tube in the millimeter region. The projected weight of such a tube is speculative. In the past, each time a new tube has been introduced at a higher frequency, the weight has initially been high but has been reduced by ultimate development. If serious development in the 35 GHz range should ever begin, the present focusing structure and therefore the overall tube weights will undoubtedly be lowered. However, it is doubtful that a tube capable of 1 kW CW at 35 GHz will ever be as light as that predicted for the microwave region. A weight of 25 lb is a more conservative estimate.

The cooling system, power supply, and prime power source for the transmitter also add to the system weight. Until adequate techniques are developed, weight estimates for the cooling system are speculative. However, it is likely that the weight of the cooling system will be insignificant compared to the weight of the required prime power source. The power supply weight should be close to that predicted by Figure 23. Even the variation in actual tube weights from that predicted in Figure 20 will be small compared to the prime power. Thus the weight required for the millimeter transmitter system should be very close to the weights calculated for the microwave system (shown in Figures 24, 25, and 26).

1.1.5 Gain

The high-power tubes discussed thus far in this section are amplifiers. For a specified power output, the gain of the amplifier determines the power requirement of the driver stage.

The gain of a TWT is a function of its length – the higher the gain the longer the tube. As the length increases beyond a certain length, focusing with a permanent magnet becomes very difficult and a solenoid must be used. With today's technology a solenoid is required if the gain of a 35 GHz amplifier is to exceed 20 dB. This technology is not expected to change appreciably in the future.

A solenoid for a 1 kW, 35 GHz tube would require a drive power of about 1 kW and additional capacity would be required in the cooling system. Therefore, the overall system efficiency would be reduced.

If a permanent magnet is used and the gain is limited to 20 dB, a 10 watt driver stage is required. The development of a 10 watt solid state driver at 35 GHz would be useful. An alternative to the solid-state driver is the extended interaction klystron oscillator.[1,13] It is capable of long life operation at 35 GHz with 10 watts (or slightly greater) output power and an efficiency of approximately 10 percent. Weights of existing tubes are approximately 5 lb.[1]

1.2 Solid-State Devices

Solid-state sources were considered in some detail in Chapter 1, Section 2. The conclusion reached was that high power solid-state devices could not be expected to compete with microwave tubes because of their poor efficiencies even if they were to match the power output (a doubtful assumption). This is also true at millimeter wave frequencies.

In the previous section of this chapter a requirement for a 10 watt solid-state source at 35 GHz was noted. This power might be obtained with a single LSA oscillator.[14] A second possibility might be paralleling or phase locking IMPATT devices.[15,16] Although the devices or techniques are not available to accomplish either of the two alternatives, it is likely that they will be in five years. The weight of the resulting device will be small enough to be neglected in the transmitter system and the power required to operate them will be approximately 100 watts.

1.3 Conclusions

Table 15 gives a summary of the performance characteristics for a potential deep space millimeter wave amplifier. The tube used is a coupled-cavity TWT with permanent magnet focusing. Its power capability is limited by the assumed capacity of the cooling system. The weight, efficiency, and lifetime represent goals which can be obtained with concentrated effort.

Also required to complete the transmitter is a driver stage capable of an output power of 10 watts. This goal is likely to be reached by solid-state sources within five years of normal development time.

The overall system weight associated with the transmitter is approximately equal to the weight of a microwave system with the same overall efficiency and output power.

Table 15

CHARACTERISTICS OF TUBES FOR
DEEP SPACE COMMUNICATIONS
AT 35 GHz

Lifetime	15,000 hours
Power	1 kW
Efficiency	40 percent
Weight	25 lb

2. ATMOSPHERIC EFFECTS ON MILLIMETER WAVE PROPAGATION

2.1 Attenuation and Sky Noise

The steady advance in millimeter wave technology suggests the possibility of planetary space probe communications using a millimeter wavelength system. It is well known that there exist two transmission windows (at 35 and 94 GHz) through the atmosphere in the millimeter wave region. But, the imperfections of these two windows must be closely examined. Comparison between relative merits of 35 and 94 GHz is also of great interest.

The total attenuation and sky temperature of millimeter waves in clear weather have been shown in Figures 34 and 35 (Chapter 1) for centimeter waves for both typical midlatitude atmosphere and dry atmosphere. Measured values available at 35,[17-19] 69.75, and 94[20] GHz are shown to be consistent with the calculated curves.

To avoid excessive absorption, very low elevation angles have to be excluded. Assuming a minimum elevation angle of 30 degrees, the estimated attenuation in clear sky for 35 and 94 GHz may be taken to be 0.5 dB and 2.4 dB from the curve $\psi = 60$ degrees in Figure 34.

Since the water drops in cloud and fog are small in size compared with a millimeter, the attenuation of millimeter waves by cloud and fog is simply proportional to their liquid water content. The cloud attenuation coefficients may be computed from the temperature and frequency dependent complex refractive index[21] and have been graphically presented for 0°C and 20°C in Figure 38 (Chapter 1). The worst temperature-zone cloud cover of long duration associated with a frontal zone is equivalent to about 1(gm-km)/m^3. The predicted attenuations for this possibly extreme condition at a 60 degree zenith angle are 2 dB for 35 GHz and 9 dB for 94 GHz. A more moderate case based on the model in Figure 39 results in an attenuation of 0.8 and 3.6 dB for 35 and 94 GHz, respectively.

Attenuation by rain is the main obstacle to millimeter wave communication. The calculated rain attenuation coefficients using Laws-Parson drop size distributions are shown in Figure 62 for the two frequencies under discussion. (These drop size distributions still are not known to be the correct ones, but they nevertheless will allow a comparison of attenuation at different millimeter frequencies to be made.) These two curves are obtained from interpolation of the data in the literature.[22,23] Space diversity might be necessary to overcome heavy rain. However, if it is essential to keep the system working in light rain, say 0.1 in/hr (which will not be exceeded for more than 1 percent of the time in most temperate regions), and assuming an average rain height of 3 km, the estimated 0.1 in/hr rain attenuations at 30 degrees elevation angle are 3 dB for 35 GHz and 11 dB for 94 GHz. The comparison between 35 and 94 is shown in Table 16.

Table 16

PROPAGATION LOSS OF MILLIMETER WAVES AT 60 DEGREES FROM ZENITH

		35 GHz	94 GHz
Attenuation through clear atmosphere (dB)		0.5	2.4
Attenuation due to dense clouds (dB)	Worst case	2	9
	Moderate case	0.8	3.6
Estimated attenuation due to 0.1 in/hr rain (dB)		3	11

For a millimeter wave system to be profitable relative to a centimeter wave system, it is essential that very dry, high-altitude sites for ground stations be selected. The favorable atmospheric conditions would minimize the adverse factors of millimeter waves. Since the loss tangent of ice is negligibly small at millimeter wavelengths, dry snow is essentially transparent to longer millimeter waves such as 35 GHz. This prediction has been confirmed by measurements at BTL.[24,25] On the other hand, the dimensions of snow flakes are comparable to the wavelengths of shorter millimeter waves such as 94 GHz, so scattering loss may occur. While reliable theoretical prediction of attenuation by snow is hardly possible, future experimental data will tell whether significant scattering loss may occur at shorter millimeter waves. This observation also implies that a ground station in an arctic or antarctic region would operate well at 35 GHz.

The sky temperatures shown in Figure 35 also apply to the millimeter waves. The values computed are for a typical clear, summer atmosphere in the temperate zone. The temperatures under clouds or light rain are higher than those for clear sky.

2.2 Error in Prediction of Refraction

The effect of atmospheric refraction may be divided into a largely predictable ray-bending component and an unpredictable short-term fluctuation component by atmospheric turbulence. Very little experimental information on atmospheric refraction is available at millimeter wavelengths. Fortunately, this refractive effect is essentially frequency independent down to millimeter waves and therefore the experience with centimeter waves should be applicable here, especially for 35 GHz. On this basis the estimated residual error in the prediction of ray bending and the rapid random fluctuations are less than about 10^{-4} rad (see Chapter 1, Section 4.2).

2.3 Phasefront Distortion

Phasefront distortion and amplitude scintillation of signals occur when electromagnetic radiation propagates

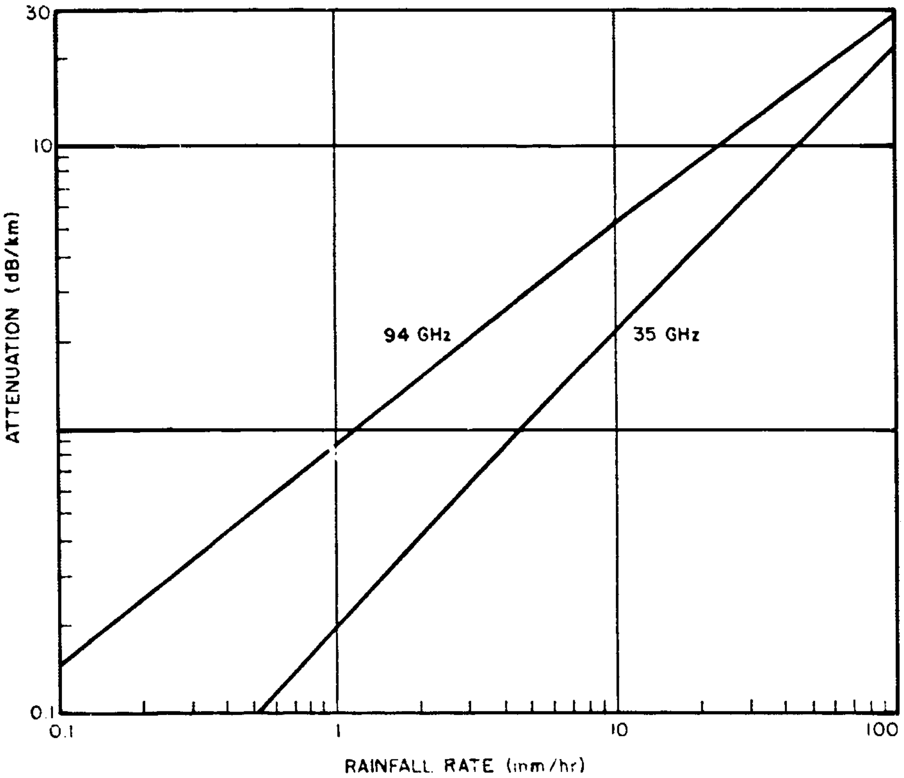


Figure 62. Rain attenuation coefficients (18°C)

through any medium that has a randomly varying inhomogenous refractive index. Amplitude scintillations are small and are usually neglected. However, in some circumstances, they may cause serious effects. These effects are considered in Chapter 4, Section 2.3. Phasefront distortions will be considered in detail here because they can seriously limit the performance of millimeter communication systems.

Signals originating from deep space sources pass through three interfering media – interplanetary space, the ionosphere, and the troposphere. The interplanetary medium contains solar plasma ejected from the Sun at high speeds. The ionosphere consists of an ionized medium in which the electron density varies with space and time. The troposphere has a randomly varying inhomogeneous refractive index caused principally by water vapor irregularities attributed to turbulent movements or random tropospheric winds.

Results of experiments[26] and theoretical work[27] at microwave frequencies have indicated that the interplanetary medium has little effect on propagation in the ecliptic plane. Millimeter signals should be affected even less by the interplanetary medium since, in general, low density plasma has less effect at higher frequencies.[28]

The ionosphere, a weakly ionized medium extending from about 50 km to about 1000 km above the surface of the Earth, can distort a signal passing through it at low frequencies, with lesser effects observed at L band and the lower portion of S band. Ionospheric effects have not placed significant limitations on communications at microwave frequencies at S band and above, and should have no measurable influence on millimeter waves.

The troposphere is the lower portion of the Earth's atmosphere and extends from the surface of the Earth to about 10 km. In addition to the attenuation, noise, and refractive effects mentioned earlier in this chapter, the atmosphere can introduce significant random fluctuations in the phase of transmitted millimeter wave signals.[29,30]

Phasefront distortion of signals passing through the troposphere is caused by random variations of the refractive index. Average values of the parameters in the equation for refractive index can be determined as a function of altitude and measured meteorological conditions, thus allowing construction of tropospheric models of major geographic and climatic regions of the world. Several of these models have been used to calculate the effect of random variations of the refractive index on signals propagated through the troposphere.[29,30] These models are useful because average conditions are represented, thus allowing estimates of phase distortion that may be typical for a given area. The extreme variability of weather with space and time limits the application of estimates of phase distortion. For example, snow, rain, and hail may occur at different altitudes simultaneously. Also, the water vapor content of the atmosphere may change with space and time, and the type of clouds and their distribution may vary. Good models for clouds have not been developed; therefore, most calculations are based on the assumption of minimal cloud cover.

Three statistical parameters are of particular interest in characterizing phasefront distortion: the spatial correlation distance, the time variation of distortion, and the rms phase deviation. The spatial correlation distance has been measured and is believed to be 100 to 200 meters depending on location and weather conditions.[31,32] The phase variation at two points on the phasefront separated by a distance equal to the correlation distance is essentially uncorrelated. The correlation time of phasefront distortion is believed to vary from several seconds for high elevation angles to several minutes for low elevation angles.[30] The importance of correlation information will be shown in later sections of this report.

The amount of tropospherically produced phasefront distortion that occurs in a signal originating from a deep space source can at least be estimated from the results of experiments conducted over ground-based paths. Theoretical estimates of phasefront distortion can also be obtained from appropriate models of the atmosphere. Although much of the experimental work has been done at microwave frequencies, the results should apply to the lower millimeter frequencies. This extrapolation should be valid to about 100 GHz, because the tropospheric refractive index can be considered frequency independent below that frequency.[31,33]

Figures 63 and 64 show theoretical results obtained by Smith[30] using model parameters based on the results of several meteorological and propagation experiments conducted by independent groups under differing weather conditions at different locations. For the purposes of this report and in accordance with the previous assumptions, the results have been extrapolated from 20 GHz to 100 GHz. The curves were truncated at 5 GHz to eliminate ionospheric effects that occur at lower frequencies. Figure 63 shows the single-path standard deviation in signal phase, with elevation angle as a parameter, that would be expected when measuring the electrical path length of many parallel paths over an area having linear dimensions of several hundred meters. Figure 64 shows the differential phase deviation that would be expected over paths separated by the indicated baseline distances. Baseline is used here to mean the line between two points located in a plane perpendicular to the direction of propagation. Figures 65 and 66 show the same results as Figure 63 and 64 except that they are for elevation angles of 15 and 30 degrees, respectively, and were obtained by using elevation angle data from Figure 63 and baseline data from Figure 64.

Two important trends may be noted in Figures 63 to 66. Figure 63 shows that the elevation angle at which the system must operate is very important. For example, distortion at 15 degrees is about twice as bad as at 90 degrees and about 4 times as bad at 0 degree as at 15

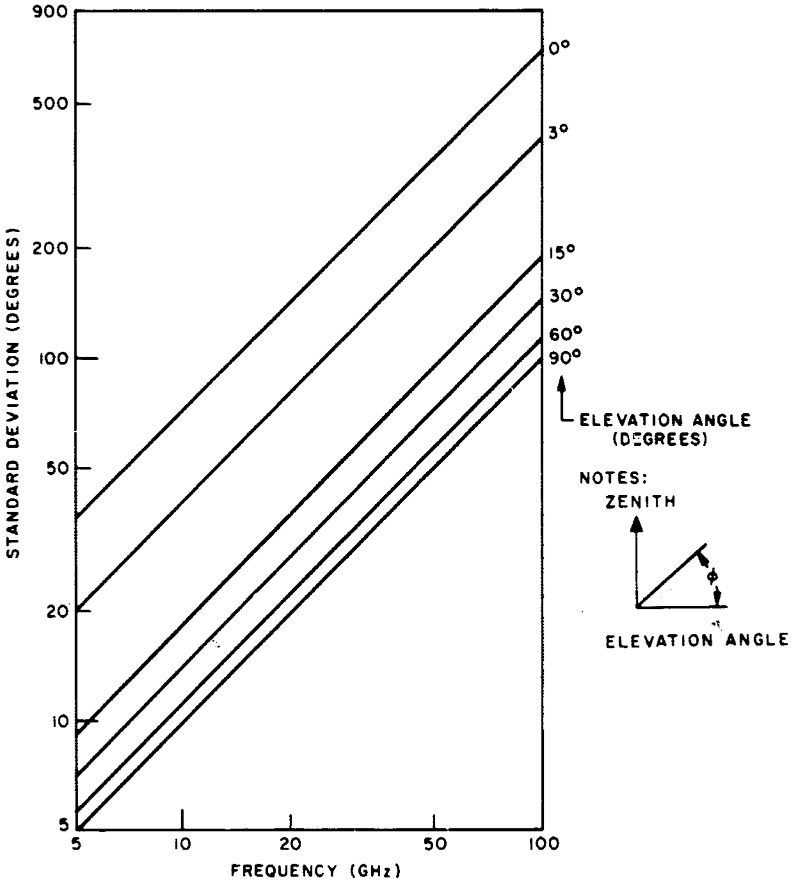


Figure 63. Standard deviation of signal phase over signal path for several elevation angles

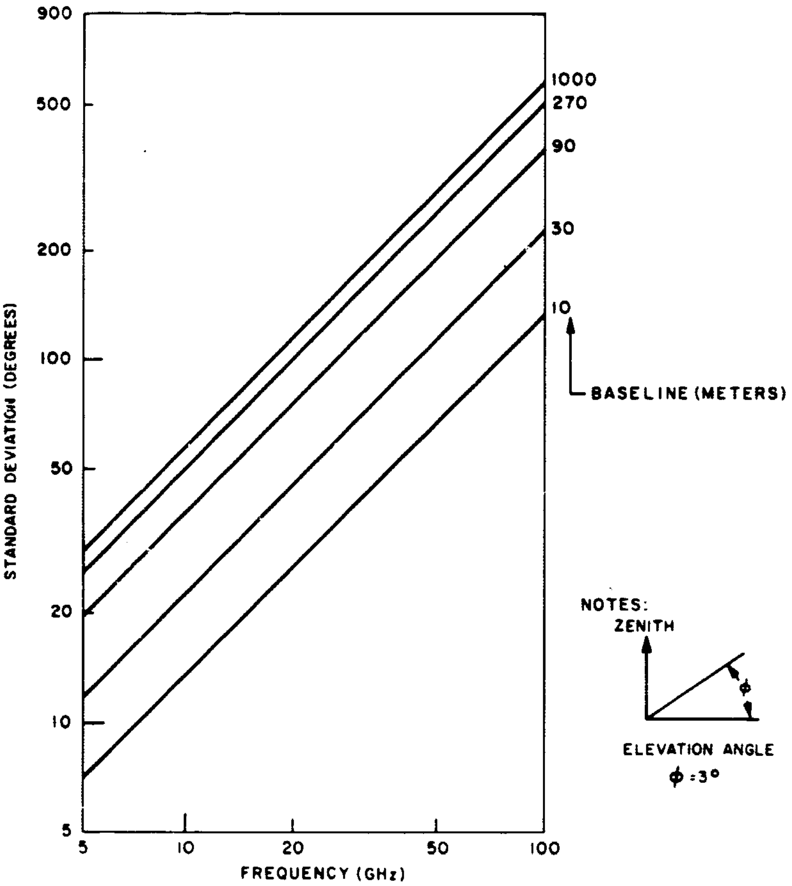


Figure 64. Standard deviation of phase difference at two receivers separated by the indicated baseline for 3-degree elevation angle

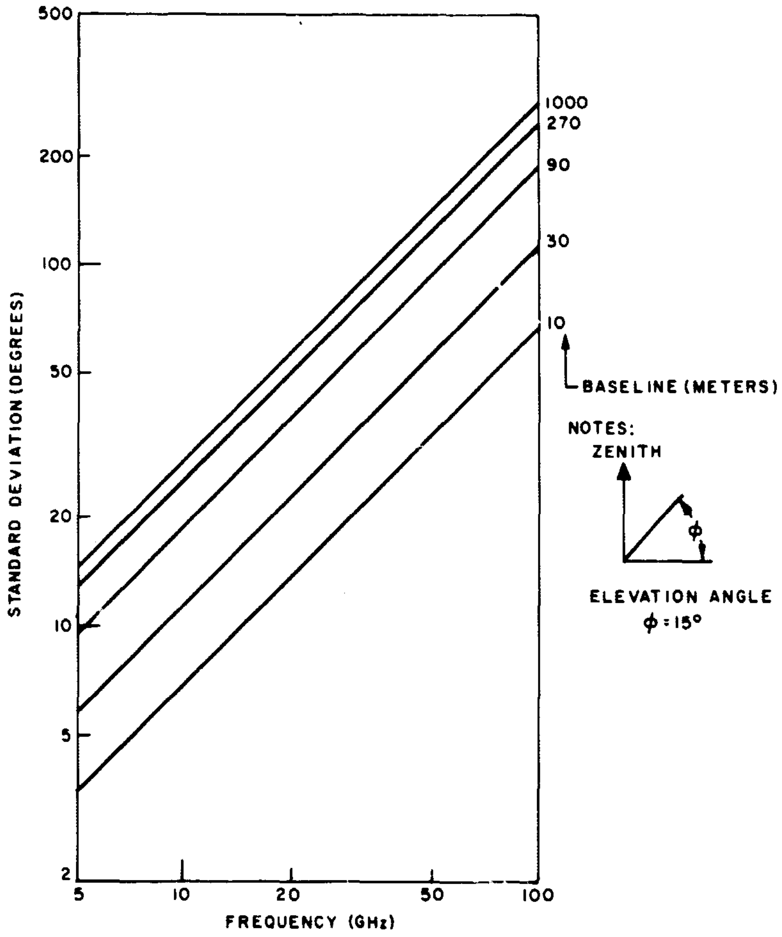


Figure 65. Standard deviation of phase difference at two receivers separated by the indicated baseline for 15-degree elevation angle

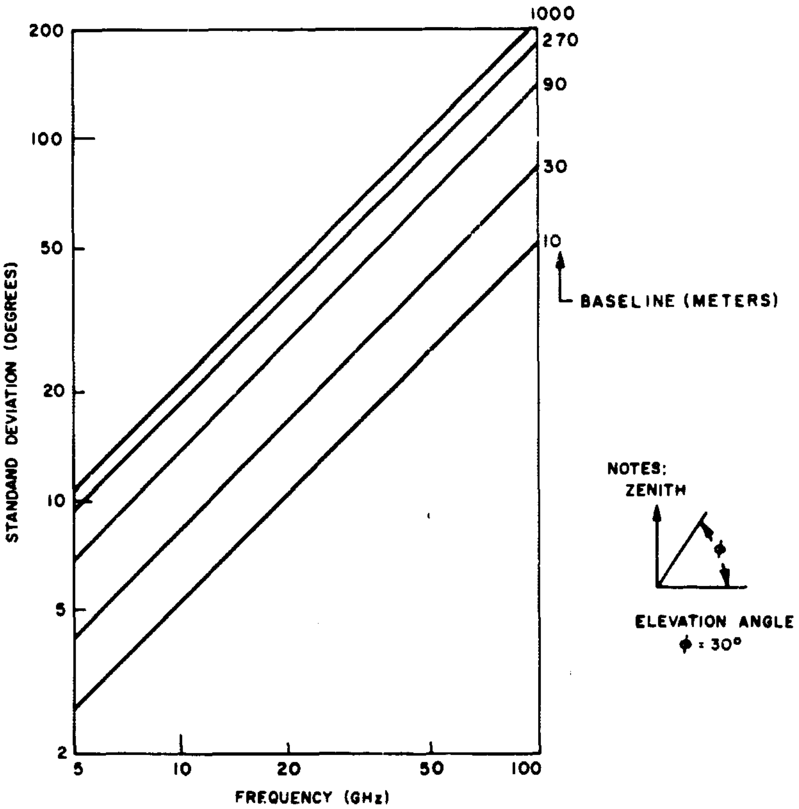


Figure 66. Standard deviation of phase difference at two receivers separated by indicated baseline for 30-degree elevation angle

degrees. Figures 64, 65, and 66 show that, for any elevation angle, parallel path phase differences increase rapidly as the baseline increases up to about 200 meters. Beyond 200 meters, phase variations are not strongly correlated and rms phase differences increase slowly.

It should be noted that all curves presented in this section represent moderately cloudy, midlatitude U.S. conditions. Clearly, the curves are average in nature and should not be interpreted as applying exactly to any particular area or time. They are useful because they can be used to obtain the magnitude and effect of phasefront distortion for "average" conditions. For example, distortion would be less severe in relatively high and dry locations such as the desert mountains of the Southwest than in the relatively high humid areas of the Southeast.

3. RECEIVERS

At present, receivers at microwave and millimeter frequencies can be separated logically by application for spacecraft or for ground station. The performances of the receivers in these categories, as well as the types of receivers, are sharply delineated. The primary purpose of this section is to summarize receiver performance for ground station (masers and parametric amplifiers); however, the subject of spacecraft receivers is also briefly considered.

Because the low noise amplifiers are only part of the receiving system, the future amplifier improvement must be considered from a tradeoff or economic standpoint in terms of system performance. The noise temperature of the receiving system is a useful measure of the overall performance. Therefore, present and projected system noise temperature will be obtained (for the ground station application) from 1 to 100 GHz.

The best spacecraft receivers employ low-noise preamplifiers such as a TWT, tunnel diode, or transistor amplifier at the lower microwave frequencies. At millimeter frequencies and higher microwave frequencies, down converters using Schottky-barrier diodes have shown promise as the most sensitive means of detection which does not involve cooling to low temperatures. The ultimate noise temperature for these types of receivers will probably be about 100°K at the microwave frequencies.

Present spacecraft have adequate up-link performance because of the high-power ground transmitters that are available. Therefore, development of improved cooled spacecraft receivers is practically nonexistent. However, in the future it may prove economical to increase the system bandwidth by increasing the frequency and by using a synchronous satellite as a repeater station for deep space communications. Under these conditions, it is probable that cooled amplifiers would be important. The traveling-wave maser as a device is well suited for remote operation; however, the associated cryogenic system reliability must be greatly improved. The problems and associated costs of overcoming them are not small, but the investment might be worthwhile. Further development of this device could also have important side effects as it would drastically reduce the maintenance costs of ground station receivers.

3.1 Devices

Since the introduction of the maser, development up to frequencies as high as 35 GHz has progressed rapidly to the point where little decrease in noise temperature can be expected in the future. Table 17 lists the performance of masers actually operating in systems today. It is anticipated that a noise temperature of at least 10°K is ultimately possible for all the masers listed in the table. In addition to AIL maser[34] at 35 GHz (listed in Table 17), several other masers at this frequency have been reported in the literature.[35,36]

Masers at higher frequencies (70, 81.3, and 96 GHz) also exist.[37-39] The noise temperatures of these masers have not been reported; however, it has been estimated that noise temperatures are in the order of 200 to 300°K.[40] Their bandwidths are in the range of only 1 to 6 MHz. Wide-band traveling-wave masers at these frequencies would be difficult to develop because of severe tolerance requirements.

The parametric amplifier has been an important competitor of the maser. Improvement in the paramp performance has been continuous until today it is approaching theoretical limits at the lower microwave frequencies.

Figure 67 shows data points which represent existing cooled parametric amplifiers. The lower curve[41] represents the theoretical limits by fully optimizing the present designs and cooling the amplifiers to 20°K. Cooling to this low temperature presents problems in the design of varactors and low-loss circulators. This will cause difficulties in reaching the projected values, particularly at the millimeter frequencies. One point of Figure 67, that at 35 GHz, is worth mentioning because it represents a recently reported[42] working model at this important millimeter frequency and because it is clearly representative of what could be expected at 35 GHz in the light of the current state of the art. It is cooled to 77°K and is voltage tunable from 34 to 38 GHz. A 94 GHz parametric amplifier is also being studied[43] by TRG, Incorporated. The major problem is the realization of an adequate varactor diode.

Table 17

MASER NOISE TEMPERATURE (°K)

System	Frequency (GHz)	Present	Future
JP-Venus[57]	2.3	5°	5°
BTL-Telstar[58]	4.16	5°	5°
JPL-Venus[59]	8.448	18°	5°
AIL[34]	35	20°	10°

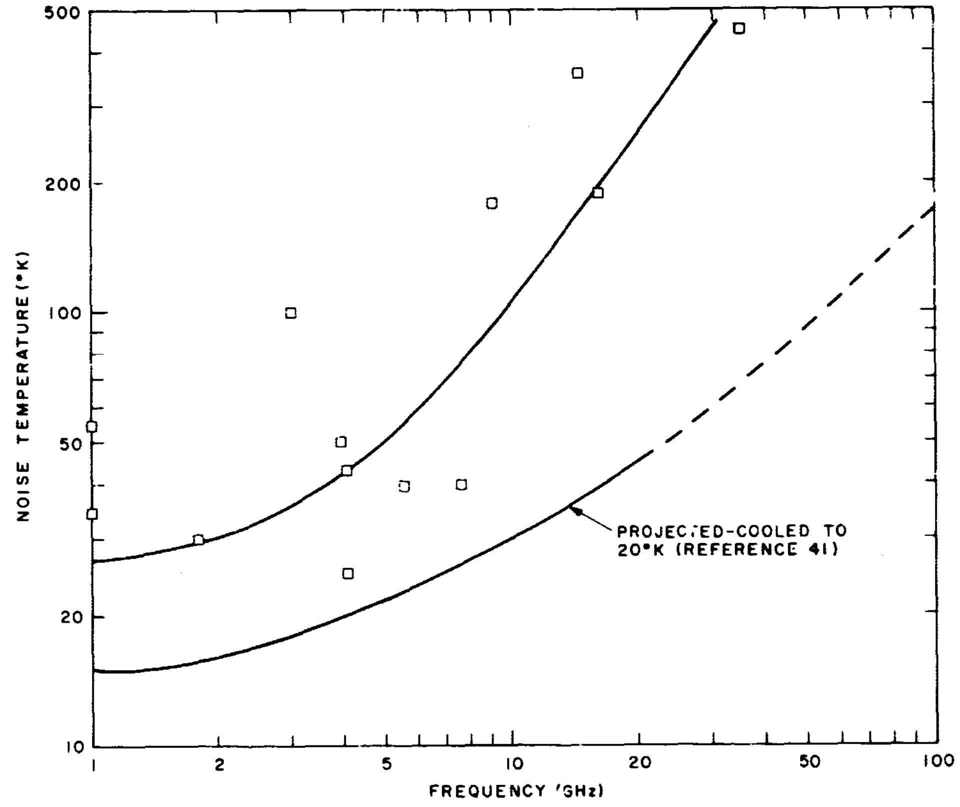


Figure 67. State of the art of cooled paramps and projected performance

No significant improvement can be expected up to frequencies of 10 to 12 GHz in the way of cooled paramps. However, paramps will continue to improve at higher frequencies as the varactor diodes are improved. Performance of varactor diodes can be measured by a commonly used figure of merit -- the cut-off frequency. For best results, the cut-off frequency of the diode should be at least ten times the pump frequency. Recently used varactors have cut off frequencies of 300 to 500 GHz.[44] However, considerable improvement has recently been obtained. Cut-off frequencies of 1000 to 1500 GHz are being obtained regularly with planar diffused gallium arsenide varactors.[45-47] Schottky barrier diodes have also been made with cutoff frequencies greater than 2000 GHz.[45,48,49] These improved diodes provide the critical element required for parametric devices as high as 100 GHz. Without minimizing the problems of designing the required coupling circuits (i.e., low-loss circulators, etc.), it is likely they can be overcome in the foreseeable future.

The paramp may be designed to operate in either a cooled or uncooled mode. In addition, it can operate continuously during the temperature variation from the cooled to the uncooled state (with corresponding system degradation). Thus it can still operate despite a breakdown in the cryogenic system. This property distinguishes it from the maser, which can operate only near cryogenic temperatures. In the uncooled mode, the paramp is the most sensitive uncooled amplifier available at microwave frequencies.

The theoretical limitation for the noise temperature of uncooled paramps using the cut-off frequency as a parameter is plotted in Figure 68. Also plotted is the state of the art (in March 1966) and the theoretical limitation associated with these paramps. It can be seen that the limit is approached but that degradation exists because of circuit losses and other noise contributed by imperfect diodes.[50]

The recently developed diodes are being used in development of laboratory-type down-converters at 32 and 94 GHz.[45] Noise figures equivalent to a single sideband noise figure of 8 to 9 dB at 32 GHz and 13 to 15 dB at approximately 94 GHz have been measured.* In addition, the diodes have been used for a down-converter at 50 GHz with a resulting noise figure of 9 to 10 dB (including the 3 to 4 dB noise figure for the i-f amplifier).[51] A noise figure of 8 dB should be obtained shortly at 50 GHz. Additional measurements[49] have been reported and indicate the possibility of 5 to 5.5 dB noise figures at 50 to 60 GHz. Others[43] have obtained a noise figure of 7 to 8 dB at 35 GHz with a down-converter; however, the diodes used do not represent the highest quality presently available. Barber has predicted an ultimate noise figure of 3 dB at X band (assuming an i-f amplifier having a 2 dB noise figure). In Figure 69, these points are plotted, in terms of noise temperature vs. frequency, along with the projection of the noise temperature which will be obtained in the next ten years.

A problem in the design of good down-converters has been the pump or local oscillator supply. Recently the LSA device[52] has been used as a local oscillator and its noise and stability performances exceed those of an available klystron at 50 GHz.[51] This device offers great possibilities for an entirely solid-state mixer at millimeter frequencies. Several efforts in this direction exist. As an example, a project at Texas Instruments is currently investigating the possibility of a 94 GHz integrated receiver on a gallium arsenide chip.[53]

3.2 Systems

In a low-noise receiver system there are contributing factors to the system noise temperature other than that of the receiver itself. Provided that system losses are small, the system temperature may be expressed by:

$$T_{sys} = T_{sky} + T_{rad} + T_{ant} + T_{rec}$$

where T_{sky} is the sky noise which consists of the tropospheric noise and the extraterrestrial noise; T_{rad} is the radome noise due to lossy material and reflections from the ground; T_{ant} is the antenna noise due to feed spillover, scattering, and resistive loss; and T_{rec} is the receiver noise which comes from the receiver temperature and the transmission line loss.

In Section 2 of this chapter and in Chapter 1, Section 4, tropospheric effects are considered. The sky noise was found to be a function of weather (clouds and precipitation), elevation angle, frequency, and location.

At frequencies as high as 35 GHz, where the maser can be used, the basic limitations in addition to the tropospheric noise are the maser temperature of 4°K and the extraterrestrial noise of 3.5°K.[54] This latter radiation is identified as isotropic in space and must be treated as a constant in the sky noise.

A few existing low-noise receiving systems[55,56] are listed in Table 18. The horn reflector antenna at Crawford Hill presents the lowest system temperature. If enough efforts are made to minimize the waveguide loss, the system temperature of a horn-reflector antenna may be further reduced to about 12°K, which is close to the achievable limit. The near-field Cassegrain antenna without a radome also has the potential to approach this limit. Although the antennas described are used at the low microwave frequencies, similar antenna noise temperature could be expected at higher frequencies.

*These measurements are currently being made. The data are tentative.

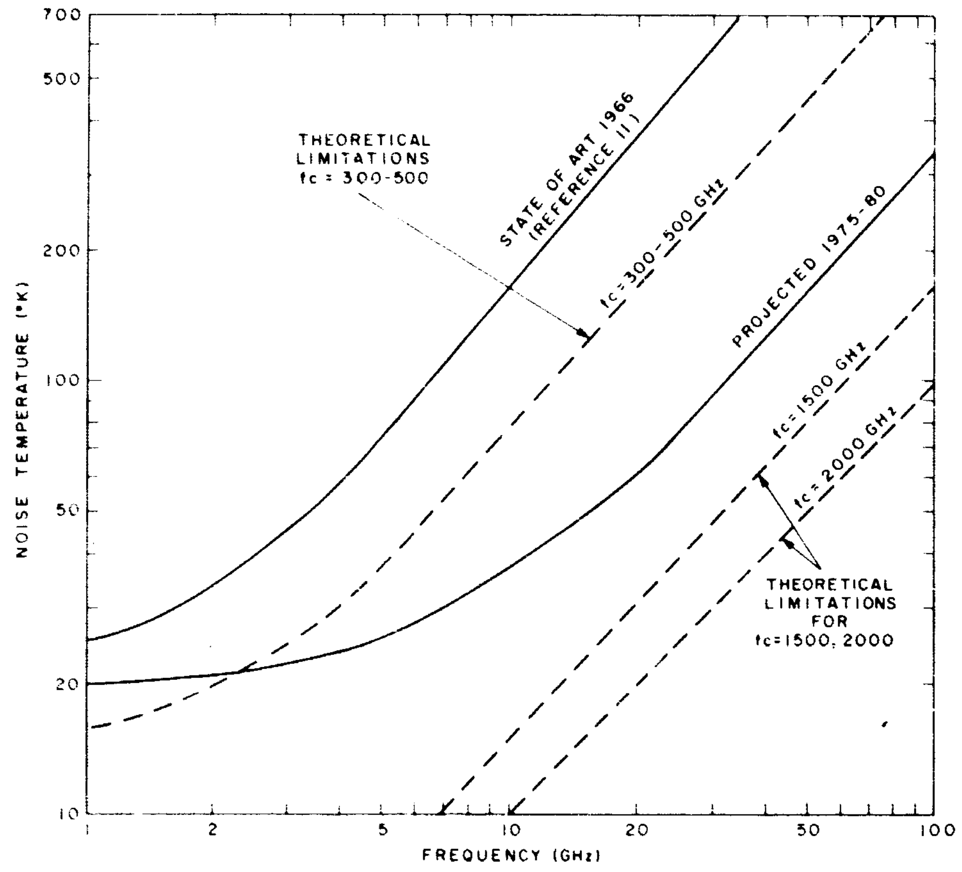


Figure 68. Noncooled paramps – state of the art and limitations based on diode cut-off frequencies

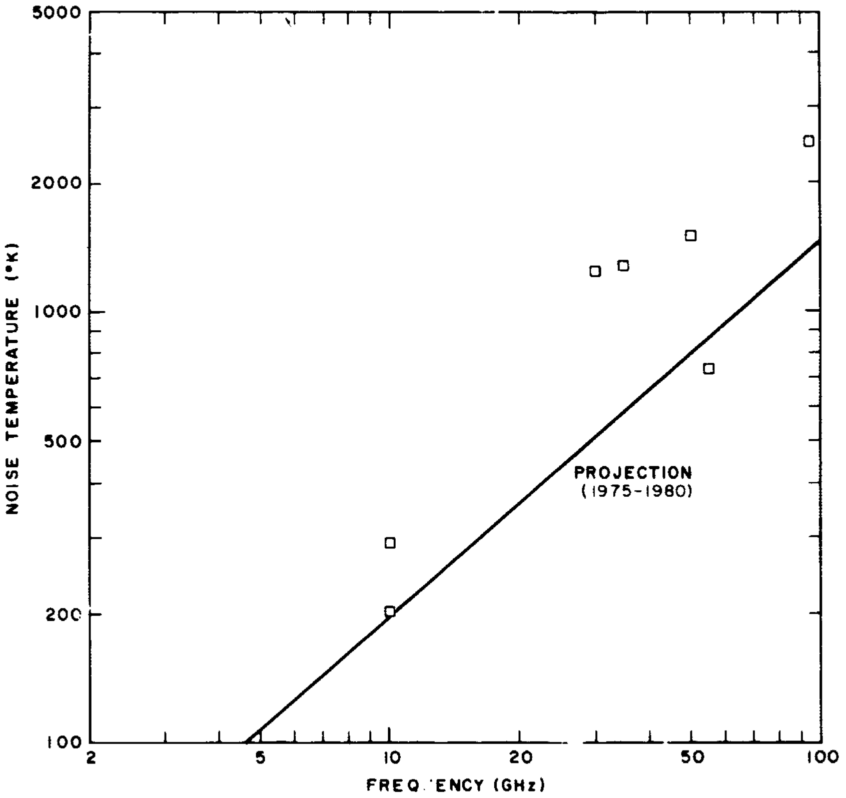


Figure 69. Noise temperature of down converters (actually obtained and projected)

Waveguide and other circuitry losses are not separately listed in Table 18; however, a factor of approximately 10°K is contributed to the system temperature of each of these systems. As frequency increases, these losses will increase proportional to $\sqrt{f}$ (at least). For the purposes of this report, this loss is assumed to contribute 15°K to the system temperature at 35 GHz.

The minimum system temperature discussed is in the zenith direction of the clear atmosphere. The tropospheric noise becomes increasingly important when the frequency increases. The curves of Figure 35 in Chapter 1 show the increase in sky temperature as a function of frequency and zenith angle. By using these data and other data listed in this section for antennas and various types of receivers, Figures 70, 71, and 72 were obtained. They show the system noise temperature from 1 to 100 GHz at zenith angles of 0 degrees, 60 degrees, and 80 degrees under clear atmosphere conditions. In the case of rain or clouds, the system temperature is increased.

A few points are worth discussion. The cost of operating at a zenith angle greater than 80 degrees at frequencies above 10 GHz is very high in terms of noise temperature. To compare systems at zenith angles greater than 60° to 70 degrees, one must be certain that such angles are indeed necessary during critical communication periods. Present systems seldom operate at zenith angles greater than 65 degrees. Therefore it may not be fair to compare system performance at these angles when such a drastic penalty is paid over some of the frequency spectrum.

If, however, the specifications logically require that such large zenith angles be considered, the comparison of noise temperatures for systems using cooled paramps and masers is interesting. The ultimate limit on the sensitivity of a ground-based receiver becomes increasingly dominated by the tropospheric effects as the frequency increases and/or as the zenith angle increases. Because of this the receiver contributes only a small part to the system performance. Thus an increase in receiver performance has a relatively small overall effect. This is most apparent at 35 GHz with a maser as a receiver where a 3 dB increase in maser performance results in less than 1 dB improvement in system performance.

4. COMMUNICATION PERFORMANCE

In earlier sections of this chapter, atmospheric effects have been discussed. These atmospheric effects require further interpretation before useful statements can be made about system performance. This section will present the effects of atmospheric turbulence, attenuation, and sky noise for the millimeter wave case and show what modifications must be made before microwave equations and concepts can be used at millimeter frequencies.

4.1 Atmospheric Turbulence

Phasefront distortion of signals passing through the atmosphere must be considered when large antenna systems

Table 18

LOW-NOISE RECEIVING SYSTEMS

	Freq (MHz)	Gain (dB)	Efficiency (percent)	Sky Noise at Zenith (°K)	Radome Noise (°K)	Antenna Temp (°K)	System Temp (°K)
NASA-JPL 85 ft diameter Cassegrain at Goldstone, California	2400	54.4	65	6	0	4.5	28
	2400	54	55	6	0	9	--
	8448	--	--	7	0	7	38
Deutsche Bundepost near-field Cassegrain at Raisting, Germany	4160	58.4	55	6.5	6	2	29
	6300	61.5	55	--	-	-	--
A.T.T.-Comsat conical horn reflector at Andover, Maine	4080	57.7	76	6	10	0.05*	34
BTL horn reflector at Crawford Hill, Holmdel, New Jersey	2390	43.3	76	6.5	0	1*	22
	4080	47.65	69	6.8	0	0.2*	20
NASA-JPL 210 ft diameter Cassegrain at Goldstone, California	2400	62	75	6	0	3.4	21

*Revised estimates based upon more recent measurements showed these less than 0.2°K.

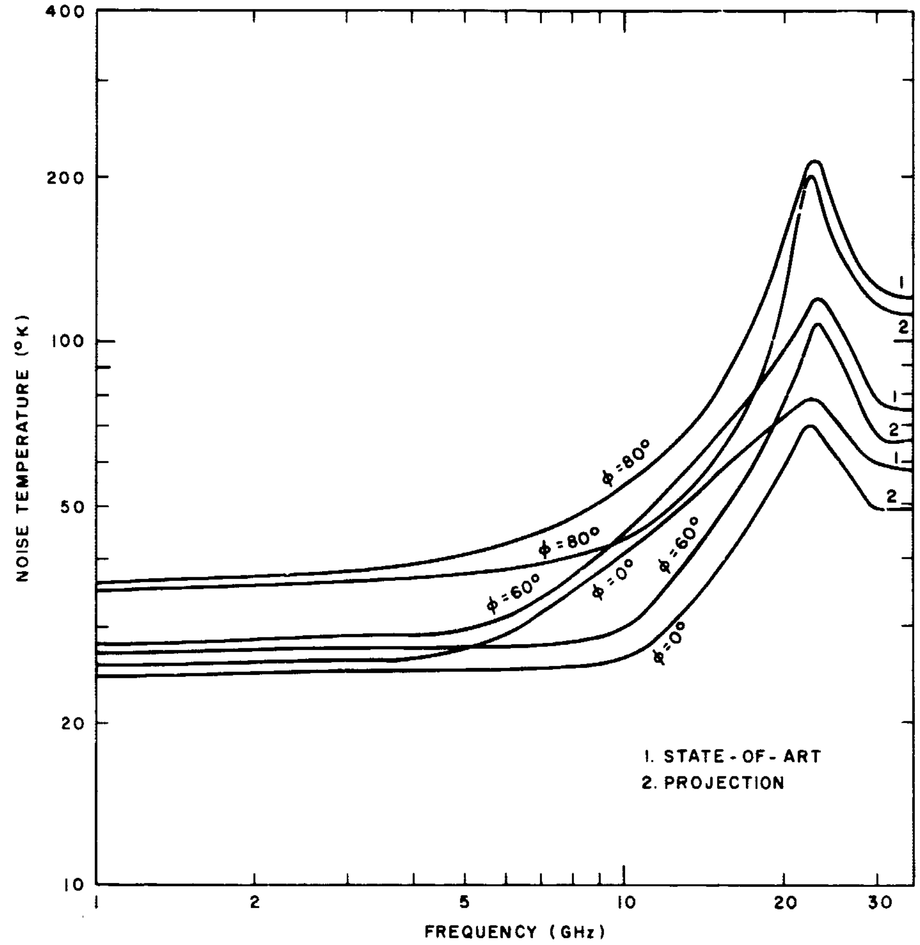


Figure 70. System noise temperatures using state of the art and projected masers at varying zenith angles

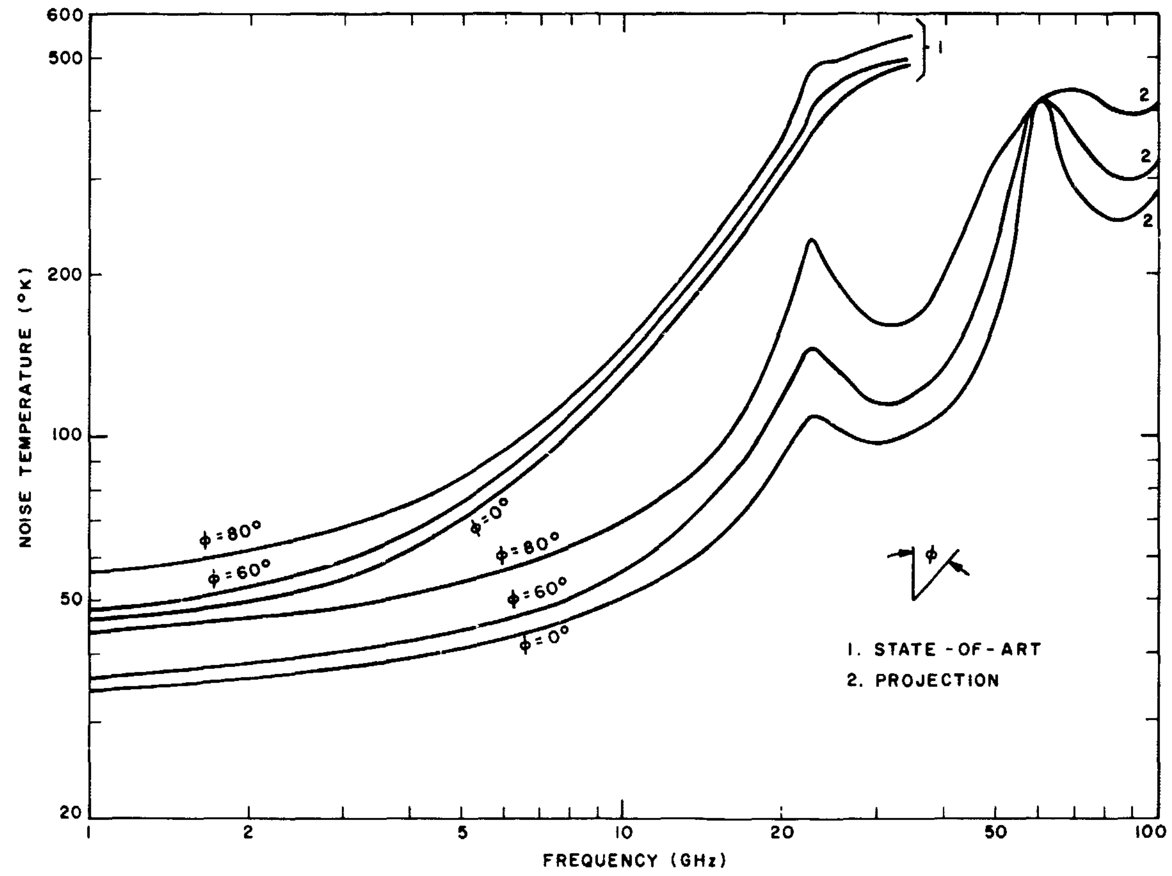


Figure 71. System noise temperatures using state of the art and projected cooled parametric amplifiers at varying zenith angles

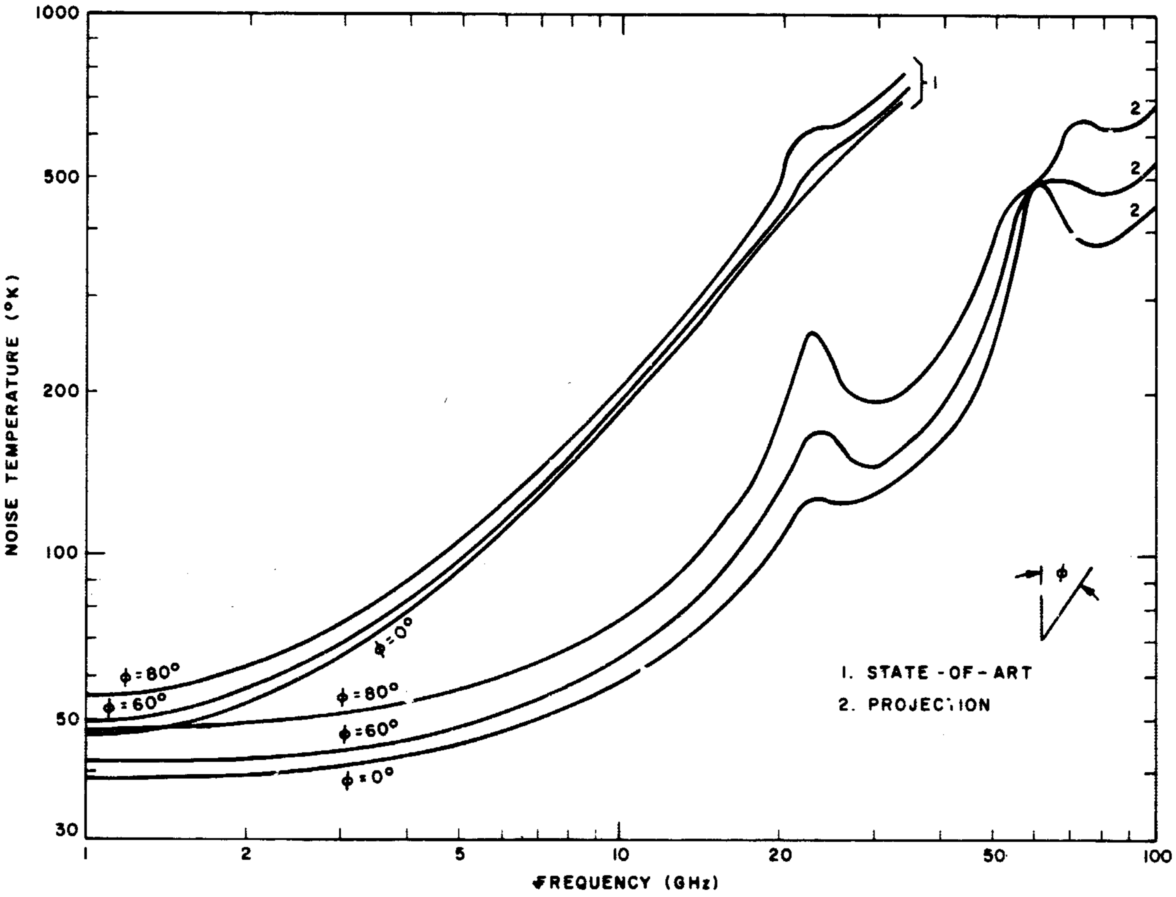


Figure 72. System noise temperatures using state of the art and projected noncooled parameters amplifiers at varying zenith angles

(i.e., both single and multiple apertures) are used. For single aperture antennas the rms deviation from a plane phasefront and the spatial correlation distance of distortion place upper limits on antenna diameter for a given frequency. For multiple-aperture antennas the time variation of distortion determines the lower limits of thresholds for the phase-tracking loops employed in dynamic correction of phasefront distortion. For simplicity, multiple aperture antennas (arrays) will be considered in Appendix 2.

The correlation distance is probably about 100 to 200 meters. Therefore, it seems that for efficient antennas the diameter should not exceed 100 meters. If the antenna diameter is larger than the correlation distance, strong correlation between phasefront fluctuations at different portions of the antenna will not exist, and destructive interference of the field will reduce the antenna gain. Large increases in aperture will result in small gain increases and reduced efficiency.

The magnitude of phasefront fluctuations across the antenna aperture will limit the antenna diameter for a given frequency. The remainder of this section will establish this bound within the limits of existing knowledge. The following assumptions will be made to simplify the procedure:

1. Antennas discussed will be parabolic reflectors with no surface errors. (This allows the bound to be established by atmospheric effects only.)
2. Deviations in the phasefront of an incoming planewave having rms value σ cause the same gain reduction as an rms error $\sigma/2$ in the reflector surface.

Antenna gain can be expressed as

$$G = \eta \left(\frac{\pi D}{\lambda}\right)^2 \quad \exp -(4\,\pi\,\sigma\,/\,\lambda)^2$$

where η is the aperture efficiency, D is the diameter, and λ is the operating wavelength. If σ is a small fraction of a wavelength, very little gain reduction will result from phase distortion. For example, if $\sigma = \lambda/32$, the gain reduction will be less than 1 dB.

Figure 73 shows the data of Figures 64, 65, and 66 expressed as a function of baseline rather than frequency. The standard deviation of the phase difference between two points separated by a distance d on a plane perpendicular to the signal direction is expressed as a function of d. Curves are shown for three elevation angles and frequencies of 35 GHz and 94 GHz. Three trends are clear:

1. Phase differences increase as d increases.
2. Phase differences increase as the elevation angle decreases.
3. Phase differences increase as frequency increases.

It is important to estimate the maximum useful antenna diameter available at millimeter frequencies. The following approach will be taken. Let the rms signal phase difference at two points on the antenna surface be less than or equal to $\lambda/6$ when the two points are chosen diametrically opposite on the minus 20 dB aperture illumination circle. For all points separated by distances less than the diameter of this aperture illumination circle, the rms signal phase difference will be less than $\lambda/6$. Thus the equivalent rms phase error for the entire antenna will be much less than $\lambda/12$.

No standard exists by which the maximum useful diameter from atmospheric turbulence considerations can be established. Any practical criterion may be chosen. It has been stated[60] that a good choice of operating frequency is in the range of 0.6 to 0.7 times the gain-limit frequency.This corresponds to an aperture rms phase error of $\lambda/21$ and results in a gain of 1.8 dB less than that obtained at the gain limit frequency. Using the $\lambda/6$ criterion, it can be shown that the frequency of operation will be below the gain-limit frequency, but not so far below that the resulting diameter limitation is overly restrictive.

Table 19 shows the aperture illuminations chosen as typical. Table 20 shows the antenna diameter at which the -20 dB points incur a phase difference of $\lambda/6$ for elevation angles of 3, 15, and 30 degrees, at 35 and 94 GHz. According to the criterion chosen, these diameters are significant because they represent basic limitations. It can also be seen that the diameter limitation is about the same for both aperture illuminations.

A particularly interesting result is the fact that, under similar atmospheric conditions, less antenna gain is available

Table 19

TYPICAL APERTURE ILLUMINATION DISTRIBUTION

Illumination Type	X at -20 dB Point	Sidelobe Level (dB)
1. Uniform $A(X) = 1$	1	17.6
2. Parabolic $A(X) = 1 - X^2$	0.955	24.6

Z = Distance off axis

Z_o = Radius

$$X = \frac{Z}{Z_o}$$

Table 20

MAXIMUM ALLOWABLE ANTENNA DIAMETER (Meters)

	Elevation Angle (degrees)					
Aperture Illumination	3	15	30	3	15	30
Uniform	16	76	170	2.4	8.3	15
Parabolic	16.8	79.5	178	2.50	8.7	15.7
	35 GHz			94 GHz		

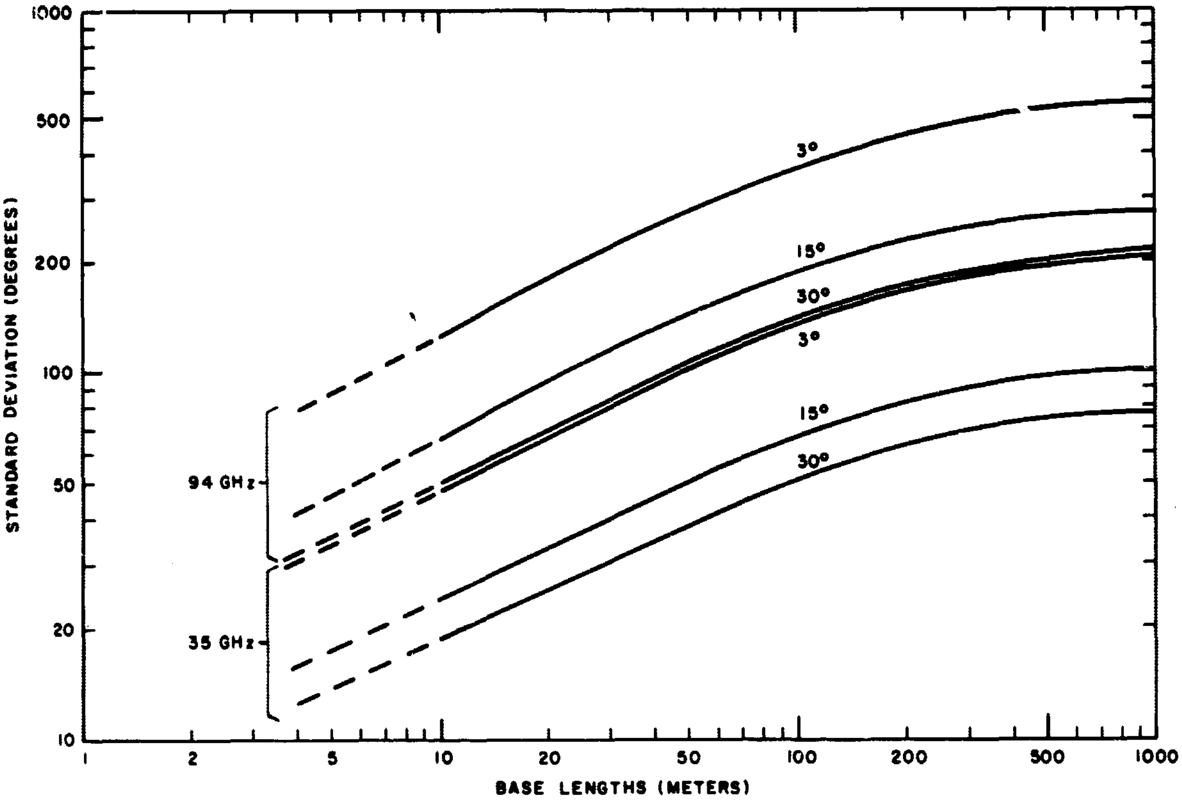


Figure 73. Standard deviation of phase difference at two receivers separated by baseline for various elevation angles

at 94 GHz than at 35 GHz. The gain of the antenna is proportional to $(D/\lambda)^2$. The results of Table 20 show that $(D/\lambda)^2$ is significantly greater for the maximum allowable antenna diameter at 35 GHz than for the one at 94 GHz. Thus, when antenna gain is limited by atmospheric turbulence, higher gains can be obtained at lower frequencies. This result represents a significant change from microwave concepts, where atmospheric turbulence presents little or no problem.[60]

Another interesting result is the effect of elevation angle on maximum antenna diameter. It is clear from Table 20 that elevation angles below 15 degrees are costly in terms of antenna gain and that elevation angles below 3 degrees may be prohibitive at millimeter frequencies. Although it has been shown that elevation angles below 25 to 35 degrees may not be necessary for deep space communications,[61] maximum antenna diameters are chosen on the basis of a 15 degree minimum allowable elevation angle.

Thus 75 meter and 8 meter antenna diameters are chosen as the largest practical size for communicating with deep space probes at frequencies of 35 GHz and 94 GHz, respectively. At least one other report[62] has concluded that 4.5 meters is the largest practical antenna diameter at 94 GHz. However, a criterion of λ/30 was chosen for the phase difference between points on the opposite edges of a 15 ft diameter antenna at 94 GHz. The resulting diameters were smaller than those obtained here.

This same report[62] used an approach in evaluating available propagation data which also is quite pessimistic. The data (from Maui, Hawaii) have a variance about four times as great as similar data from the mainland United States.[63] To avoid pessimistic outcomes, the data must be considered along with an appropriate model to account for meteorological conditions, as shown by Smith.[63] The Maui data were taken at an elevation angle of 6 degrees. Data at higher elevation angles are not available at present; however, the use of the Maui data without correction to a higher elevation angle is also too pessimistic (at millimeter frequencies in particular).

Millimeter frequency antennas on the order of 75 meters and 8 meters for 35 GHz and 94 GHz have not been built and may never be built because of structural "cost effective" arguments set forth in Chapter 1, Section 5. It can be concluded that turbulence will not limit communications except at very low elevation angles, at the highest millimeter frequencies, and under conditions of unusually turbulent weather.

4.2 Atmospheric Attenuation

Electromagnetic energy propagated through the atmosphere can be attenuated by water vapor, oxygen, rain, fog, and ice. The amount of attenuation depends on several meteorological parameters as well as on signal frequency. Atmospheric attenuation for microwave frequencies was discussed for several weather conditions in Chapter 1, Section 4 and for millimeter frequencies in Section 2 of this chapter. It is clear from both sections that operation at millimeter frequencies requires that more system margin be allowed for atmospheric attenuation. This section will discuss methods of determining such margins and will attempt to estimate the margins. Examples are given in Table 21. Clear-sky, cloud, and rain attenuation will be considered separately.

Table 21

MARGIN REQUIREMENTS FOR SYSTEM TO OPERATE AT 70° ZENITH ANGLE IN THE BOSTON AREA WITH 5 PERCENT OUTAGE TIME BECAUSE OF WEATHER

Type of Attenuation (dB)	Frequency (GHz)			
	6	15	35	94
Clear sky	0.08	0.20	0.75	3.6
Cloud and Rain	<0.08	0.5	2.3	≈7
Total	<0.16	0.7	3.05	≈10.6

Clear-sky attenuation at a given frequency (attenuation by oxygen and water vapor) depends upon absolute humidity, elevation, and elevation angle. Clear-sky attenuation is always present and can be estimated for a given frequency, elevation angle, and location. When an estimate is made of margin required to overcome attenuation due to rain and clouds, clear-sky attenuation will be added.

Cloud attenuation varies with cloud density and path length through the cloud. These factors are required to calculate cloud attenuation from Figure 38. Since total overcast occurs at least several percent of the time in most locations,[64,65] it is necessary to obtain a measure of how large the water content may be. Preliminary measurements of Sun-signal attenuation at 30 GHz made at BTL suggest that excess attenuation is small during a high percentage of overcast periods. These measurements were made at zenith angles greater than 60 degrees. The results further suggest that the "integrated" water density may be small for a large percentage of total overcast conditions.

Rain is the most significant attenuator at microwave and millimeter frequencies; therefore, rain attenuation will be discussed in some detail. As documented in Chapter 1, Section 4, there is no detailed correlation between the rain rate measured on the ground and sky attenuation (or correspondingly sky noise). Comparison of system operation at different frequencies may still be made, using attenuations based on Laws-Parsons distributions.

As an example, it can be seen (from Figure 62) that if the rainfall rate were 16mm/hr, signal attenuation would be as follows:

Frequency (GHz)	Attenuation (dB/km)
3	(4.5×10^{-3})
35	4
94	8

If rain at this rate were uniform and perpetual, communication at millimeter frequencies would be difficult. Conversely, if rain never fell, operation at millimeter frequencies might be attractive. An approach more useful than the previous comparison is to determine the probability that atmospheric attenuation due to rain will exceed some value, A, as a function of zenith angle and frequency. Comparisons may then be made. The attenuation distributions for the Boston area, shown in Figures 45 and 46, may be used at 15 and 35 GHz. The frequency-dependence of rain attenuation and the frequency relationship for cloud attenuation in Figure 38 permit reasonable estimates of attenuations at frequencies other than 15 and 35 GHz. The data for 6 and 94 GHz in Table 21 were obtained in such a manner for a zenith angle of 70 degrees and a 5 percent outage time.

These values are only rough estimates. This is partly due to the coarseness of the data of Figures 45 and 46 at low time percentages. Values of attenuation at small time percentages (< 1 percent) may be required when considering deep space communications. To obtain these values, additional measurements covering more periods of heavy rain are required.

The Boston area is not an ideal site for space communications because of the relatively high annual rainfall. Presumably, at better locations such attenuations would correspond to a smaller outage than 5 percent. However, the attenuation differences at the frequencies shown in the table are indicative of the problems that exist for low elevation angle operation at 35 and 94 GHz as compared to 6 GHz.

Continuous operation at millimeter frequencies may be impractical. If so, an allowable outage time must be specified for millimeter system operation. A system margin requirement is then defined as the maximum atmospheric attenuation encountered for the given allowable outage time. By determining the margin requirements for various locations, performance comparisons can be made. Extensive data and good statistics for attenuation distributions at preferred locations would be most helpful.

4.3 Sky Noise

The increase in effective antenna temperature because of reradiation of energy by atmospheric gases and precipitation reduces system sensitivity. It is desirable, therefore, to be able to predict the fraction of time that sky noise exceeds chosen levels. With that knowledge, enough system margin could be allowed to overcome the noise a large percentage of the time. Oxygen and water vapor, clouds, and rain contribute to sky noise, and an approach similar to that in the previous section is taken. By using the known relationship between sky noise temperature and attenuation, distributions for attenuation described in Section 4.2 can be used to obtain the noise temperature margin requirements.

Sky-noise temperatures were estimated for the conditions in the example of the previous section. The results, in Table 22, show maximum sky temperatures for a system in the Boston area for 95 percent of the time, at an antenna elevation angle of 20 degrees.

Table 22

SKY NOISE FOR SAME OPERATING CONDITIONS IMPOSED IN TABLE 21

Frequency (GHz)	6	15	35	94
Total sky Noise Temp (°K)	10.5	43.5	147	265

4.4 Conclusions

The effects of atmospheric turbulence, attenuation, and noise have been discussed in this section. The results indicate that millimeter systems should be located in the best possible locations from a weather standpoint and perhaps rely on diversity of ground stations if a high probability of uninterrupted communication is desired. Microwave systems appear to be rather immune to weather-produced outages except possibly for extreme cases. The system implications of atmospheric effects on microwave and millimeter systems are discussed in Chapter 5.

REFERENCES

1. D. C. Forster, "High Power Sources at Millimeter Wavelengths," Proc. IEEE, Vol. 54 (April, 1966), pp 532-539.

2. B. Kulke, Millimeter-Wave Generation With Electron-Beam Devices – A Critical Survey of the State-of-the-Art, NASA TN D-3727 (February, 1967). See also B. Kulke and C. M. Veronda, "Millimeter-Wave Generation with Electron-Beam Devices," Microwave Journal. Vol. 10 (September, 1967), pp 45-53.

3. Parametric Analysis of Microwave and Laser Systems for Communication and Tracking, Phase I Final Report, Contract No. NAS-5-9637, Report No. P66-16, Hughes Aircraft Company (February, 1966).

4. Millimeter Communication Propagation Program, First Quarterly Report, Contract No. NAS5-9523, Raytheon Company, Space and Information Systems Division (February, 1965).

5. Study of High Power Transmitting Tubes Above X-Band, Final Report, USAE-COM Contract No. DA-28-043 AMC-00234(E) (July 1965).

6. D. C. Forster, "Mid-1964 Review of Available Millimeter-Wave Sources," Wescon Technical Papers, Vol. 8 (1964) Part 5, presented August 25 – 28, 1964.

7. D. C. Forster, K. H. Fuller, J. F. Heney, et al, Research on Millimeter-Wave Tubes, Hughes Research Laboratories, Technical Report AFAL-TR-65-282 (October, 1965).

8. J. F. Heney, "Some New Results With High Power Millimeter-Wave Tubes," Wescon Technical Papers, Vol. 8 (1964) Part 5, presented August 25 – 28, 1964.

9. M. J. Schindler, "Can Traveling Wave Tubes Be Scaled?" Microwave Journal, Vol. 9 (September, 1966), pp 43-47.

10. Private communication with J. F. Heney, Hughes Research Labs, Malibu, California.

11. Applied Research on Efficiency Improvement in O-type Traveling Wave Tubes, Final Report, Contract No. AF33(615)-1951 (April, 1967).

12. M. N. Ernstoff and O. G. Sauseng, Advancements in TWT Efficiency With Combined Voltage Jump, Velocity Tapers and Enhanced Beam Bunching, presented at the International Electron Devices Meeting, October, 1966.

13. W. R. Day, "The Millimeter-Wave Extended Interaction Oscillator," Proc. IEEE, Vol. 54 (April, 1966), pp 539-543.

14. Private communication with John Copeland, Bell Telephone Laboratories, Murray Hill, New Jersey.

15. C. B. Swan, T. Misawa, Louis Marinaccio, "Composite Avalanche Diode Structures for Increased Power Capability," IEEE Trans., Vol. ED-14 (September, 1967), pp 584-589.

16. H. Fukui, "Frequency Locking and Modulation of Microwave Silicon Avalanche Oscillator Diodes," Proc. IEEE (Letters), Vol. 55 (March, 1967), pp 451-452.

17. K. N. Wulfsberg, "Sky Noise Measurements at MM Wavelengths," Proc. IEEE, Vol. 52 (March, 1964), pp 321-322.

18. K. N. Wulfsberg, "Atmospheric Attenuation at Millimeter Wavelengths," Radio Science, Vol. 2 (March, 1967), pp 319-324.

19. D. H. Ring, Final Report, AF 19 (122) – 458 to Lincoln Lab, Bell Telephone Laboratories, 1956.

20. A. W. Straiton and J. C. W. Tolbert, "Factors Affecting Earth-Satellite Millimeter Wavelength Communications," IEEE Trans., Vol. MTT-11 (September, 1963), pp 296-301.

21. D.E. Kerr, Propagation of Short Radio Waves, (New York, McGraw-Hill Book Company, 1951), p 675.

22. R. L. Mitchell, Radar Meteorology at Millimeter Wavelength, Aerospace Corporation, El Segundo, California, U. S. Air Force Contract No. AF04(695)-669.

23. R. G. Medhurst, "Rainfall Attenuation of Centimeter Waves, Comparison of Theory and Measurement," IEEE Trans., Vol. AP-13 (July, 1965), pp 550-564.

24. R. W. Wilson, Millimeter Wave Sun Tracker, presented to 1968 USNC/URSI-IEEE, April, 1968, Washington, D. C.

25. Final Report on Microwave Research, Rep No. 4269-K (March, 1956), Bell Telephone Laboratories, Contract No. AF-19(122)-458.

26. Mahlon Easterling and Richard Goldstein, "The Interplanetary Medium and S-Band Telecommunications," Astronautics and Aeronautics (August, 1966), pp 80-86.

27. N. L. Kaydanovskiy, N. A. Smirnova, "Resolution Limits of Radio Telescopes and Radio Interferometers Imposed by Propagation of Waves in Space and in the Atmosphere of the Earth," Radio Engineering and Electronics, Vol. 10 (September, 1965), pp 1355-1362.

28. Millimeter Communication Propagation Program, Final Report, Vol. I, Raytheon Company, NAS5-9523 (November, 1965).

29. B. R. Bean and G. D. Thayer, "Models of the Atmospheric Radio Refractive Index," Proc. IRE, Vol. 47 (May, 1959), pp 740-755.

30. P. G. Smith, "Atmospheric Distortion of Signals Originating from Space Sources," IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-3, No. 2 (March, 1967), pp 207-216.

31. T. S. Chu, private conversation.

32. K. A. Norton et al, An Experimental Study of Phase Variations in Line-of-Sight Microwave Transmissions, N.B.S. Monograph 33 (November 1, 1961).

33. Modern Radar, R. S. Berkowitz, editor (New York, John Wiley and Sons, Inc., 1965).

34. F. R. Arams and B. J. Peyton, "Eight-Millimeter Traveling-Wave Maser and Maser-Radiometer System," Proc. IEEE, Vo. 53 (January, 1965), pp 12-23.

35. R. Genner and W. M. Nixon, "Ferric-Doped-Rutile 8mm Maser," Electronic Letters, Vol. 2 (November, 1966), pp 406-407.

36. A. Robert and Y. DeCoatpont, "Traveling-Wave 8mm Maser," Electronic Letters, Vol. 3 (January, 1967), p 5.

37. W. E. Hughes, "Maser Operation at 96 kMc with Pump at 65 kMc," Proc. IRE (Correspondence), Vol. 50 (July, 1962), p 1691.

38. W. E. Hughes and C. R. Kremenek, "70-Gc Maser," Proc. IEEE (Correspondence), Vol. 51 (May, 1963), p 856.

39. W. E. Hughes and C. R. Kremenek, "An 81-Gc/s Zero-Field Maser," Proc. IEEE, Vol. 54 (April, 1966), pp 623-627.

40. Private communication with C. R. Kremenek, Westinghouse Electric Corporation, Baltimore, Maryland.

41. W. G. Matthei, "Recent Developments in Solid State Microwave Devices," Microwave Jourral, Vol. 9 (March, 1966), pp 39-46.

42. Electronically Tunable Parametric Amplifier, Final Report, TRG, Incorporated, RADC Contract No. AF30(602)-3829 (December, 1966).

43. Development of a Low-Noise Millimeter-Wave Parametric Amplifier, Final Report, TRG, Incorporated, USAECOM Contract No. DA36-039-AMC-02365(E) (April, 1967).

44. C. L. Cuccia et al, "RF Design of Communication-Satellite Earth Stations (Part 2)," Microwaves (June, 1967), pp 27-37.

45. Private communication with C. A. Burrus, Bell Telephone Laboratories, October 2, 1967.

46. C. A. Burrus and T. P. Lee, "A Millimeter-Wave Quadrupler and an Up-Converter Using Planar Diffused Gallium Arsenide Varactor Diodes," submitted to IEEE Trans. on Microwave Theory and Techniques for publication.

47. C. A. Burrus, "Planar Diffused Gallium Arsenide Millimeter-Wave Varactor Diodes," Proc. IEEE, Vol. 55 (June, 1967), pp 1104-1105.

48. D. Kahng, "Au-n-Type GaAs Schottky Barrier and Its Varactor Application," BSTJ, Vol. 43 (January, 1964), pp 215-224.

49. D. T. Young and J. C. Irvin, "Millimeter Frequency Conversion Using Au n-Type GaAs Schottky Barrier Epitaxial Diodes with a Novel Contacting Technique," Proc. IEEE (Correspondence) (December, 1965), pp 2130-2131.

50. M Uenohara, "Noise Consideration of the Variable Capacitance Parametric Amplifier," Proc. IRE, Vol. 48 (February, 1960), pp 169-179.

51. Private communication with W. M. Hubbard et al, Bell Telephone Laboratories.

52. J. A. Copeland, "CW Operation of LSA Oscillator Diodes – 44 to 88 GHz," BSTJ, Vol. 46 (January, 1967), pp 284-287.

53. Private communication with E. Maynard, Wright-Patterson Air Force Avionics Laboratory, concerning MM Wave Integrated Receiver Front End, Contract No. AF33(615)-5102, July 25, 1967.

54. A. A. Penzias and R. W. Wilson, "A Measurement of Excess Antenna Temperature at 4080 Mcs," Astrophys. J., Vol. 142 (July, 1965), pp 419-421.

55. D. C. Hogg, "Ground Station Antennas for Space Communication," Advances in Microwaves, Vol. 3, Academic Press, to be published.

56. C. T. Stelzried, Improved Calibration Techniques: X-Band Noise Temperature Calibrations, SPS 37-44, Vol. III (March 31, 1967), Jet Propulsion Laboratory, Pasadena, California, pp 72-85.

57. R. C. Clauss, A Traveling Wave Maser for Deep Space Communications at 2295 and 2388 MHz, TR-32-1072, Jet Propulsion Laboratory, Pasadena, California (February, 1967).

58. W. J. Tabor and J. T. Sibilia, "Masers for the Telstar Satellite Communications Experiments," BSTJ ,Vol. 42 (July, 1963), pp 1863-1886.

59. S. M. Petty and R. C. Clauss, Low Noise Receivers: Microwave Maser Development, SPS 37-42, Vol. III, Jet Propulsion Laboratory, Pasadena, California (November 30, 1966), pp 42-46.

60. P. Potter, W. Merrick, and A. Ludwig, Large Antenna Apertures and Arrays for Deep Space Communications, JPL Technical Report No. 32-848, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California (November, 1965).

61. Advanced Deep Space Communications Study, Final Report, Hughes Aircraft Company, Contract No. NAS 12-81 (January, 1967).

62. Millimeter Communication Propagation Program, Final Report, Vol. I, Raytheon Company, Contract No. NAS 5-9523 (November, 1965).

63. P. G. Smith, "Atmospheric Distortion of Signals Originating from Space Sources," Proc. IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-3 No. 2 (March, 1967), pp 207-216.

64. United States Weather Bureau, Summary of Hourly Observations, No. 82-26, Las Vegas, Nevada (1963).

65. United States Weather Bureau, Summary of Hourly Observations, No. 82-4, Bakersfield, Califo͏ınia (1963).

CHAPTER 3. OPTICAL TECHNOLOGY

1. LASERS

Two basic applications of the laser are considered for deep-space optical communications. They are the communications transmitter aboard the deep-space vehicle and an optical beacon either on Earth or on an Earth-orbiting vehicle. Discussion of transmitter lasers will be confined to CW oscillators because: (1) the bit rate objective, $\sim 10^6\ s^{-1}$, is higher than pulse relaxation rates of relevant lasers and (2) the preferred forms of modulation are binary polarization or phase shift FM (Chapter 4, Section 5). Only high-power pulsed lasers are considered for an optical beacon because of SNR and beamwidth relationships at the beacon receiver (Chapter 4, Section 4). In this section, selected laser oscillators are evaluated from the viewpoints of power, efficiency, size, weight, life, and reliability. Efficiency characteristics relative to theoretical limits and possible means for more closely approaching these limits are discussed.

1.1 Laser for Communications Transmitter

Bandwidth and SNR considerations (Chapter 4, Sections 5 through 9) require an optical output power $P_{out} \gtrsim 1$ watt for a communications transmitter at a range ~1 AU. To achieve optimum transmitter gain, which is essential for a high-performance optical system, the laser must oscillate in the diffraction-limited TEM$_{00}$ mode. Surface tolerances of the transmitting antenna mirror and beam-pointing precision required for diffraction-limited transmission are discussed in Sections 3 and 4.

Continuous-wave laser oscillators which yield an output in the watt range include: argon[1] (0.48 and 0.51μ), second harmonic generation (SHG) Nd:YAG.[2,3](0.53μ), ruby[4] (0.69μ), GaAs[5] (0.84μ), Nd:YAG[4] (1.06μ), Ho:YAG[4] (2.12μ), Dy:CaF$_2$[4] (2.36μ), and CO$_2$[1] (10.6μ). Four of these eight types, namely argon, both Nd:YAG lasers, and CO$_2$, will be discussed in depth. The others are less advantageous at present because the CW ruby laser output is unstable (spiking) and because CW oscillation of Ho:YAG, Dy:CaF$_2$, and GaAs requires low temperatures (~77°K).* Although low-temperature laser cooling is technically possible, the penalty in power efficiency and system complexity would be substantial. Also, for state-of-the-art GaAs lasers, restriction of oscillation to the TEM$_{00}$ mode generally reduces the total CW output power substantially.

Characteristics of the four leading candidates for a deep-space optical transmitter appear in Table 23. TEM$_{00}$ power levels listed in the table have been experimentally realized, with laser cavity lengths $\lesssim$1 meter and with laser pumps which appear compatible for space qualification. Higher TEM$_{00}$ power levels should be achievable, at least for Nd:YAG and CO$_2$ lasers, with longer (including folded) cavities and with comparable power efficiencies. As an example, 30 watts TEM$_{00}$ radiation is reported for a 2 meter long CO$_2$ laser.[6] The laser beam is coupled out in most cases by partial transparency of one of the two cavity mirrors. Output power in the TEM$_{00}$ mode, for a given cavity diameter and length, is maximized by appropriate selection of mirror curvature.[1]

Laser cooling, which is needed for all types discussed, is an exceedingly critical problem from at least three viewpoints. They are: (1) cooling system power, (2) cooling system weight, and (3) vibration produced by flowing coolants. The last may pose the most formidable technical problem if the requisite pointing precision is an arc second or less (see Section 4 for effects of vibration on pointing accuracy).

1.1.1 Argon Ion Laser Cavity Design

Extensive parametric studies of argon laser characteristics have been published.[1, 7 to 11] A strong consensus, based on these results and also on more recent development work of TEM$_{00}$ mode argon lasers, is that optimum performance is achieved with a bore diameter ~ 2.5 to 3 mm and a discharge length ~ 30 to 60 cm. Pump power

*L.A.D'Asaro and J.C. Dyment of BTL recently achieved CW operation at 200°K heat sink temperature, the highest reported so far for GaAs.

Table 23
LASER CHARACTERISTICS

Laser	λ (microns)	$P_0(TEM_{oo})$ (watts)	Observed Basic Efficiency* η_B (percent)	Theoretical* Efficiency η_{BT} (percent)	Length of Active Medium (cm)	Diameter of Active Medium (cm)
Argon	0.48	1.5	0.04 dc Pump	7	30-60	0.25-0.30
	0.51	1.5	0.06 rf Pump†			
Nd:YAG	0.53	>6‡	0.2‡	30	5	0.6
Nd:YAG	1.06	12	0.4	30	5	0.6
CO_2	10.6	10§	10§	40	100	1.2

* See Section 1.2 for discussion.

† Data on rf ring excitation from Reference 19.

‡ Results[3] with newly developed SHG materials show that power and efficiency at 0.53 micron will be $\geqslant$ 50 percent of 1.06 micron performance.

§ For flowing gas systems; values for static systems are $\lesssim$ 70 percent of this value (Section 1.4.2).

into the discharge is ~ 6 kW for a total TEM_{oo} beam power ~ 3 watts (see Section 1.2). A critical factor affecting optimum design is the beam confining magnet.[7 to 11] Magnet weight and size, as well as P_{in}/P_{out} ratio, decrease rapidly with decreasing diameter; on the other hand, diffraction losses in the cavity and discharge wall erosion by charged particle bombardment become excessive at very small bore diameters. The axial magnetic field is typically 1000 gauss for a 2.8 mm bore discharge tube. Pumping can be either rf[1] or dc;[7 to 11] overall power efficiency is comparable for both types of excitation (Section 1.2). References 12 and 13 cover examples of argon laser designs with simultaneous high performance in four parameters: TEM_{oo} power, power efficiency, size, and weight, all of which are important for spaceborne lasers. Brewster angle windows yielding a linearly polarized output beam and dielectrically coated external cavity mirrors are used in nearly all argon lasers. The output beam contains several wavelengths in the blue-green spectral region with a power distribution typically 45 percent at 4880Å, 45 percent at 5145Å, and 10 percent at other wavelengths.[1, 7 to 10] It also is possible to attain ~ 70 percent of P_{total} in TEM_{01} at 5145A.[13] Tunable wavelength selection can be achieved with a prism inside the laser cavity. Recent results[14] with very high current (30 to 300 A) wall-stabilized arc discharges show that high efficiencies η_B can be attained without a magnetic field in 7 to 15 mm bore argon lasers. Total input power is high, however, and no data on TEM_{oo} power appear in Reference 14.

1.1.2 Nd:YAG Laser Cavity Design

Published parametric design investigations for Nd:YAG are not as complete as for the argon types. A basic reason is that experimental variation of laser length and diameter is limited severely by available Nd:YAG crystal and pump lamp sizes. Typical laser rod diameters are 0.25 to 0.6 cm with lengths of 2.5 to 5 cm. Power output in the TEM_{oo} mode as high as 12 W for a 3 kW pump power level is achievable (present lamp life at this power is ~ 1500 hours). An ellipse is used to focus the optical pump (usually a tungsten lamp) on to the laser rod. Optimum design requires that the lamp image match the laser rod diameter. Double elliptical cavities (two lamps) provide higher TEM_{oo} power efficiency than a single ellipse (one lamp) because of pump threshold effects. Because of the low gain in Nd:YAG, ~ 2 percent per cm, high optical quality of crystal end face polish and of the dielectric mirrors are critical for peak efficiency. Most power output data reported for the 1.06 micron Nd:YAG laser are for an unpolarized output beam; i.e., end faces of the laser rod are cut perpendicular to the axis. Linearly polarized output beams can be obtained by addition of a Brewster angle plate inside the cavity, which reduces P_{out} only slightly with high-quality optical components. Polarized outputs also can be attained by cutting at least one end of the Nd:YAG rod at Brewster's angle. With this approach, however, off-axis cavity mirror alignment is required. This feature can introduce mechanical design difficulties, since cooling of the laser rod and pump is essential. Also a much longer crystal boule is needed to obtain the same output power with a Brewster angle cut as for crystals cut orthogonal to the axis. The 1.06 micron line width increases with temperature, which reduces the laser power efficiency. Preferred operating temperature is $\lesssim$ 30°C, with ~ 70°C as an upper limit. At this point substantial losses appear in efficiency from line broadening.

Efficient second harmonic generation (SHG) at 0.53 microns in Nd:YAG has only recently been achieved in the

CW mode.[3] A nonlinear crystal oriented for SHG phase matching is placed inside the cavity. The material used in earlier[2] pulsed 0.53 micron SHG work was $LiNbO_3$. Attempts for CW operation with $LiNbO_3$ have been unsuccessful, or at best extremely inefficient, because of optical inhomogeneities or damage produced in $LiNbO_3$ by intense optical radiation.[15] Recent developments of a new material, $Ba_2NaNb_5O_{15}$, which is more resistant to optical damage, has permitted a SHG power conversion efficiency of ~ 50 percent.[3] Increased conversion efficiency is expected with further experimentation (an efficiency approaching 100 percent is theoretically possible).[2] Temperature of the SHG crystal must be maintained near 80°C to within ~ 0.2°C to maintain phase match conditions.[3] Power needed for such constant temperature control is a few watts which is negligible relative to pump lamp input power.

Output power of the Nd:YAG for a fixed pump power per unit length should be at least linearly proportional to the laser crystal length. Significant increases in rod length, if longer crystals with appropriate optical quality can be grown, should be possible before diffraction or pumping losses become limiting. Accordingly, TEM_{oo} power levels higher than those given in Table 23 for ~ 5 cm rod length should be achievable with longer pumps and laser crystals. As previously mentioned, limitations on availability of Nd:YAG crystal lengths and pump lamp geometries have impeded studies of longer cavities. Since the overall Nd:YAG laser head length is small (~12 inches; see Section 1.3), compared to those of gas lasers, further development of longer Nd:YAG lasers would merit consideration in an optical space communications program.

1.1.3 CO_2 Laser Design

Conditions for optimum CO_2 laser efficiency, in terms of gas discharge conditions, gas pressure, gas additives, temperature, and other factors, have been studied comprehensively.[1, 16 to 20] Output power is relatively insensitive to discharge tube bore but depends strongly on type of gas additive(s) mixed with the CO_2 and also on the linear rate of gas flow through the laser tube. Typical conditions for optimum TEM_{oo} P_{out} with a 1 meter long discharge are: (1) 1 cm bore discharge tube, (2) CO_2:N_2:He gas mixture at respective pressures of approximately 2:1.5:6 torr, (3) ~ 20 mA discharge current at 9 kV, (4) gas flow rate $\gtrsim$ 50 cc/min STP (standard temperature and pressure), and (5) dielectrically coated cavity mirrors on a substrate transparent at 10.6 microns (e.g., Irtran II). Data[17] on gain per unit length as a function of gas flow rate, with gas mixture as a parameter, are shown in Figure 74. Experience has indicated that optimization of gain vs. flow rate and gas mixture closely coincides with optimum P_{out}.

Stoppage of gas flow (i.e., static or sealed-off lasers) from ~ 50 cc/min STP reduces P_{out} by $\gtrsim$ 30 percent. This effect is due in part to accumulation of CO (by CO_2 dissociation) in static systems and also is a critical factor affecting life (Section 1.4). It should be noted that discharge conditions stated above for optimum P_{out} in a flowing laser can be substantially different in an optimized static laser, as discussed in more detail in Section 1.4.

Cold cathodes normally are used because CO_2 laser current is low, cathode fall potential is small compared to total discharge voltage, and the oxidizing atmosphere of the CO_2 discharge tends to poison heated cathodes. A ballast impedance is required in series with the power supply and discharge. Minimum ballast power is dissipated with a constant-current electronic regulator rather than a passive resistance. Output power and gain become higher as discharge wall temperature is decreased, as illustrated in Figure 75. Thus, laser cooling to $\lesssim$ 30°C again is essential. Internal mirrors attached to the discharge tube appear advantageous because of adverse absorption losses and/or hygroscopicity of available Brewster window materials at 10.6 microns. Experience indicates that deterioration of dielectric coatings of internal mirrors by the discharge, a problem in the argon laser, is very low in CO_2 lasers.

1.2 Efficiency

Most data on laser power efficiency pertain to the basic efficiency, η_B, defined as the percent ratio of output power, P_o, to pump power dissipated in the active laser medium P_{il}. In general, P_o and η_B do not reach maximum at simultaneous experimental conditions; however, operation near maximum TEM_{oo} P_o levels usually is the best compromise between optimum power and efficiency. Consequently, in discussions of P_o and η_B values in this section, emphasis is placed on maximum achievable P_o at a laser life $\gtrsim 10^3$ hours. Experimental η_B data and theoretical maximum efficiency η_{BT} are given in Table 23. Definition of η_{BT} is the ratio E_{ul}/E_{pg}, where E_{ul} and E_{pg} are energy differences respectively of the upper to lower laser levels and the upper pump level to ground state level of the active atom or molecule. It should be noted that the relevant argon ground state is that of the neutral atom rather than the Ar^+ ion.

1.2.1 CO_2

The observed basic efficiency is closest to theoretical for CO_2, within a factor of ~ 5. The maximum multimode η_B reported for CO_2 is ~ 20 percent, or one-half of η_{BT}, which is close to the best that may be expected in the future. Static CO_2 gas lasers would require less weight (Section 1.3) on a space mission but at a penalty in both efficiency (at least a 30 percent reduction based on present experience) and life (Section 1.4). The minimum usable flow rate is ~ 10 cc/min STP (Figure 74), which is the value used for weight calculations of CO_2 gas storage in Section 1.3.3.

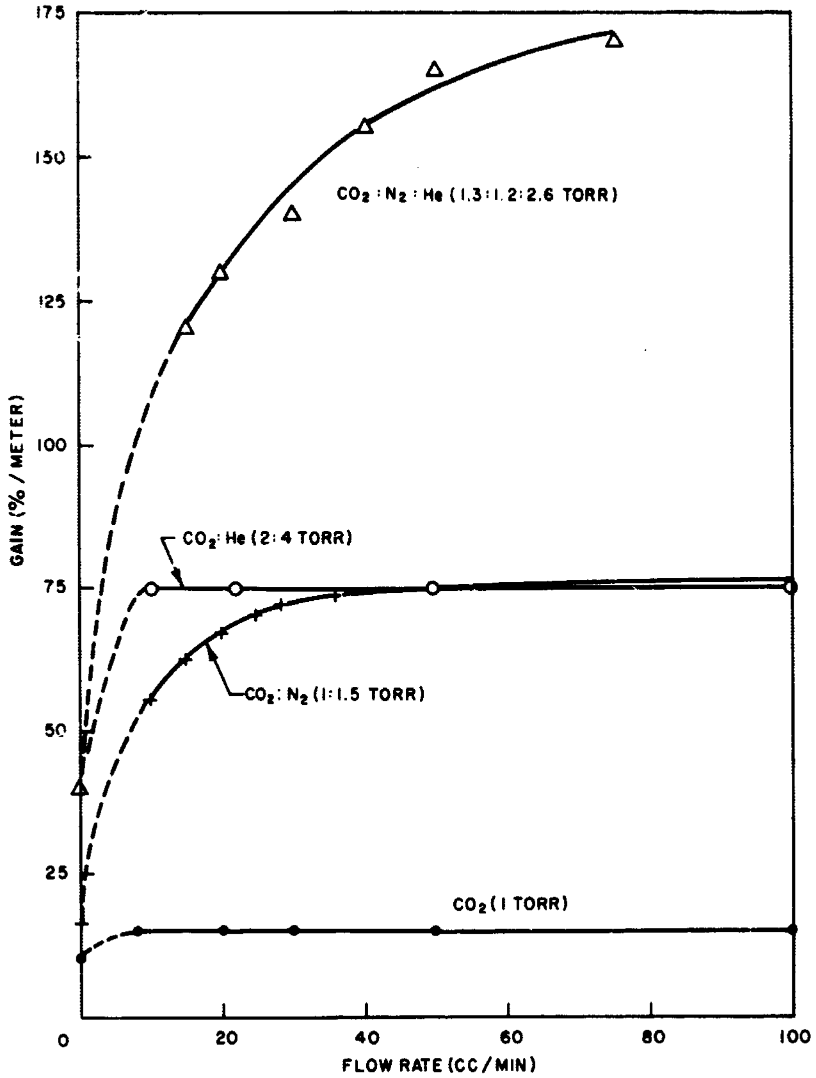


Figure 74. Gain of 22 mm bore CO_2 laser amplifiers with various gas mixtures vs. flow rate (reference 17)

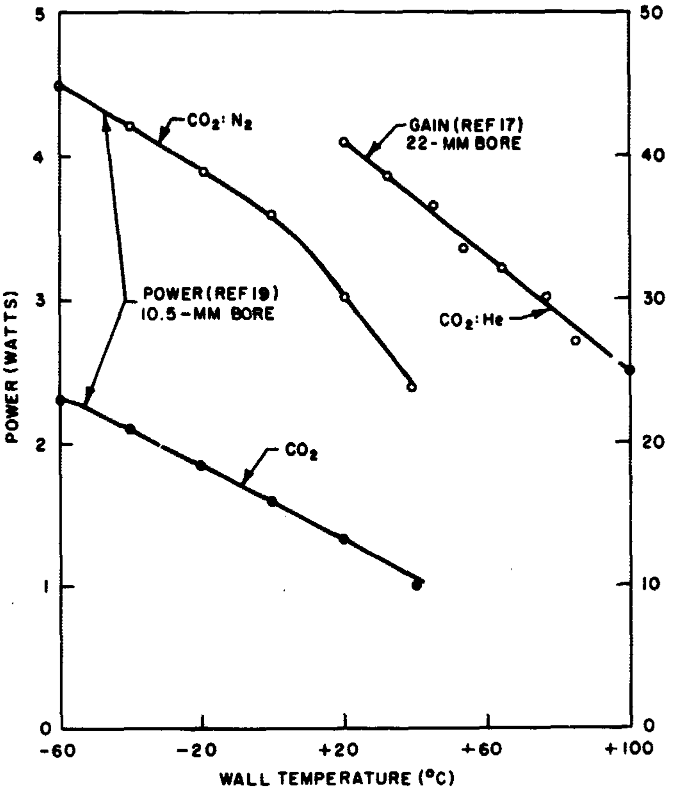


Figure 75. Dependence of output power and gain of static (nonflowing) CO_2 lasers on discharge wall temperature

1.2.2 Nd:YAG

In the remaining portion of Section 1, the 0.53 micron and 1.06 micron lasers will be treated as a single type. For the Nd:YAG laser, η_B is ~ 10^{-2} of theoretical efficiency, which is almost entirely due to a spectral mismatch between the optical lamp and the useful pump bands of Nd:YAG in the 0.5 to 0.9 micron range.[4,21] Proposed approaches for improving the spectral match include new lamp types and special doping of Nd:YAG to widen the pumping band.[4] Some points that merit mention are that any new pump lamp must have a geometry that matches the Nd:YAG rod in the lamp focusing structure, that long lamp life is essential, and that any doping added to Nd:YAG should not significantly degrade the optical quality of the crystal or broaden the 1.06 micron linewidth of the Nd dopant. Experience to date suggests that new lamp types, e.g., electroluminescent diodes, offer much greater promise than selective crystal doping.

1.2.3 Argon

The present basic efficiency of the argon laser is also ~ 10^{-2} of theoretical. The outlook for efficiency improvement in argon is enhanced by the square law dependence[7] of P_o on discharge current I. Since $P_{in} \propto I$, the basic efficiency $\eta_B \propto I$. Basically the task is to reduce tube erosion by improved discharge tube structures and cooling techniques. Possible means for achieving this include: (1) segmented metallic discharge tube,[7,11] (2) more efficient cooling of discharge tube, e.g., radiation or heat pipes (Section 1.3.4), and (3) reduction of ion loss at walls, e.g., by higher magnetic field. The last item, of course, involves a trade-off with weight. Basic efficiency of the rf-excited laser is about 1.5 that of dc-pumped lasers. This advantage is more than compensated for by a less efficient and heavier power supply, as discussed in the next section.

1.2.4 Overall Efficiency

For systems comparisons the relevant quantity is the total efficiency η_T, which includes power consumption of ancillary components such as power converters, cathodes, ballasts, magnets, and laser cooling systems. A comprehensive evaluation of η_T is premature because of the paucity of lasers specifically engineered for space or airborne application. It is of merit, however, to examine semiquantitatively the factors that contribute to η_T; these factors are summarized in Table 24. Available data on dc-to-dc power converters in the 10 V to 10 kV range indicate that ~ 90 percent conversion efficiency is feasible (Chapter 1, Section 3). For the rf-pumped argon ion laser the power conversion efficiency is ~ 50 percent.[22]

Cathode heater power is required only for the argon laser; a cold cathode is proposed for the CO_2 laser (Section 1.3). Recent cathode research[23] shows that a heated hollow cathode offers advantages, when used in high current lasers, in life and in cathode fall potential relative to oxide or matrix types. Power consumption of this cathode is ~ 25 watts for the argon laser power levels given in Table 23.

Ballast power is essential for the CO_2 laser and probably is necessary for the argon laser, depending on the output beam stability required. Minimum ballast power is dissipated in the CO_2 case with an electronic constant-current regulator. Experience indicates that a ballast dissipation of 0.1 to 0.2 of discharge excitation power is adequate (0.15 is assumed in Table 24).

Table 24

FACTORS AFFECTING TOTAL POWER EFFICIENCY η_T (TEM$_{oo}$ Mode)

Laser	Power Converter Efficiency (percent)	Cathode Power (watts)	Ballast (watts)	Magnet Power (watts)	Cooling System (watts)*	η_T -- Excluding Cooling (percent)
Argon	90dc pump 50 rf pump†	25	100‡	1000	3300	0.03
Nd:YAG	90	None	None	None	$\lesssim$560 (0.53μ) 280 (1.06μ)	$\gtrsim$0.18 (0.53μ) 0.36 (1.06μ)
CO_2	90	None (cold)	$0.15P_{in}$	None	13§	8∢

* Watts to be dissipated by the cooling system per watt output (for lasers of Table 23).

† Reference 22.

‡ Optional, depending on stability required (see Section 1.4).

§ For flowing gas system. It is $\gtrsim$18 watts for static CO_2 systems.

∢For flowing gas system. η_T is $\lesssim$5 percent for static systems.

Weight considerations presently argue against a permanent magnet for the argon laser. However, preliminary results on a periodic permanent magnet show promise for the future.[13] Required power dissipation for an appropriately designed electromagnet producing ~ 1000 gauss axial field for the argon laser dimensions given in Table 23 is not more than 900 to 1000 W and probably will decrease with advances in technology.

Power required for the cooling system is expressed as watts dissipated per watt of laser output. Quantitative determination of cooling system power (and, incidentally, weight and size) await more engineering work on heat dissipation in spaceborne laser systems (see Chapter 1, Section 2 for status of microwave and millimeter wave cooling systems).

1.3 Size and Weight

Engineering of lasers for space or airborne application is in the early stages. Consequently, availability of data on size and weight, as well as on power efficiency and life, is extremely limited compared to that on microwave tubes. The following discussion should be interpreted in that light. In many cases, it is necessary to extrapolate information on lasers developed for applications where weight and size were not of critical importance. Future improvements in size, weight, and efficiency are potentially much greater, percentage-wise, for lasers (and perhaps also millimeter wave tubes) that for microwave tubes. Table 25 summarizes size and weight estimates.

1.3.1 Argon

References 12 and 13 are examples of the present state of the art for an airborne air-cooled laser and serve as a basis for estimates given for argon in Table 25. The weight of an rf power converter would be significantly greater, by perhaps about a factor of 2, than the 50 pounds projected for a dc supply. This would argue against use of an rf argon laser unless significant advantages exist in other parameters, such as life, over the dc-pumped type.

1.3.2 Nd:YAG

Size given in Table 25 has been achieved in a commercial model (Korad Model K-Y1). A weight of 10 lb

Table 25

LASER SIZE AND WEIGHT ESTIMATES* OF LASERS IN TABLE 23

Size Excluding cooling (cu. ft.)	Argon	Nd:YAG	CO_2
Laser head	1.0 ft^3 (32″X9″X6″)	0.1 ft^3 (12″X4″X4″)	1.0 ‡ ft^3 (48″X6″X6″)
Power converter	1.0	0.2†	0.5
Weight Excluding cooling (lbs)			
Laser head	25	10	25‡
Magnet	25	None	None
Power converter	50	15†	25
Total weight	100	25†	50‡
Cooling System			
Watts dissipated/ watts P_o	3300	280	13

* See discussion in Section 1.3.

† Power converter may not be needed at all if voltage of the Nd:YAG optical pump lamp can be matched to that available from the spacecraft power supply.

‡ Does not include gas storage if a flowing gas CO_2 system is used. It would require ~ 0.7 lb of premixed gas per 1000 hours of flow at a 10 cc/min STP rate (10 grams/mol gas weight assumed, e.g., CO_2:N_2:He in 2:1:12 ratio). Volume could be conserved by storage in liquid or solid state at low temperature.

excluding cooling would seem reasonable in view of what has been accomplished for an argon head. As noted in the table, the possibility exists that Nd:YAG power could be taken directly from the spacecraft raw supply.

1.3.3 CO_2

Estimated laser head size is somewhat smaller in cross section than argon but 16 inches longer. Laser head weight of CO_2 is assumed equal to that of argon; weight of the more rugged discharge tube and anode structure necessary for the argon laser is assumed equal to weight of the additional 16 inches of CO_2 head length. Power-converter size and weight are based on an input power of ~ 125 W, vs. ~ 10 kW for argon, and include a constant-current electronic ballast dissipating 15 W.

Weight of gas storage for a flow system is rather high (Table 25, footnote ‡). The weight advantage of the static CO_2 system, however, is offset at present with a lower efficiency and life (Sections 1.2 and 1.4).

1.3.4 Comments on Cooling System Size and Weight

Although the argon laser appears to bear the greatest penalty in size and weight of the cooling system, it can operate at a much higher temperature (~ 1500°C) than the other two laser types. Direct cooling by radiation through a window in the spacecraft may then be feasible, similar to that proposed for some microwave tubes (Chapter 1, Section 2). Another cooling technique which may be promising for space systems is the heat pipe (Chapter 1, Section 2) which may have advantages in efficiency and reduction of vibration (of concern in pointing precision – Section 4).

1.4 Life and Stability

1.4.1 Life

The present status of life and the primary life-limiting factors for the three laser types are given in Table 26. The limitations listed in the table are well recognized and intensive development efforts are in progress at many laboratories to increase the life of all three types. In view of this state of flux, it is perhaps most appropriate to list the magnitudes of present life performance and the principal approaches under investigation for life improvement, which are also shown in Table 26. Within the next five years, life objectives of at least 10,000 hours for all three laser types would not appear unreasonable.

Factors that have limited life in the past, and are not included in Table 26, are deterioration of mirrors and Brewster windows. Progress has been made in understanding the relevant physical mechanisms and in developing the necessary technology so that windows and mirrors no longer are regarded as primary components that limit life.

Table 26

LIFE LIMITATIONS OF LASERS

	Argon	Nd:YAG	CO_2
Life at present (hr)	1000-2000	1500	No known limit (flowing) 300-1000 * (static)
Primary life limitation	Discharge tube structure	Pump lamp	Gas dissociation and cleanup (static)
Other life limitations	1. Cathode 2. Gas cleanup		
Means of increasing life	1. Segmented metal discharge tube 2. Hollow cathode 3. Argon reservoir and/or automatic leak valve 4. Minimize electrode sputtering	1. New lamp types 2. Development of longer life tungsten lamps	1. Electrode and discharge wall materials to minimize gas cleanup 2. Oxidizing agents to convert CO into CO_2

*Life data in the range of 3000 to 5000 hours have been reported for static CO_2 systems; however, the discharge conditions for the longer life were such that the power efficiency was much less, ~ 1-2 percent, instead of the 7 to 10 percent given in Table 23.

1.4.2 Nonflow (static) CO_2 Laser Life

The life of the static CO_2 laser became an area of intensive study during 1967 and a few comments on what has emerged are of merit. The basic life-limiting mechanism (Table 26) is dissociation of CO_2 and a resultant gas cleanup by gettering and/or precipitation.

Comparisons of gain and spectral content of visible sidelight emission, between static and flowing CO_2 discharges, show that a substantial concentration of CO accumulates in static systems within a few minutes after the discharge is activated.[17] This dissociation of CO_2 causes the excitation conditions for optimum laser output power to be substantially different from those of flowing systems with a concomitant loss in efficiency. Gas decomposition of both CO_2 and CO occurs continuously for CW discharge operation until the CO_2 partial pressure becomes insufficient to sustain laser oscillation. Experiments show that gas cleanup is greatest near the cathode electrode; two relevant processes are gettering action by metallic films sputtered from the catnode and formation of a carbonaceous deposit (possibly including precipitation of carbon) on tube walls near the cathode. Preliminary life tests on static CO_2 lasers show that laser action terminates after ~300 to 1000 hours when the laser is operated at maximum output power. Life times as high as 3000 and 5000[24,25] hours have been reported for static lasers, but their basic efficiency (~1 to 2 percent) was below optimum by a factor of about five.

The factors which are important for long-life CO_2 lasers are not fully understood. Techniques which have been proposed include:

1. Use of electrode materials such as platinum that minimize sputtering and cleanup
2. Addition of water vapor
3. Use of quartz rather than glass discharge tubes
4. Rigorous outgassing of laser structure prior to filling with CO_2 mixture
5. Addition of oxidizing agent or catalyst to decrease CO_2 or CO dissociation rates
6. Minimizing gas contamination by careful selection of all materials, including cements and mirrors, which are exposed to the discharge.

Determination of the relative importance of these parameters awaits further study.

1.4.3 Stability

Pheonomena that produce noise in the output laser beam include:

1. Plasma instabilities of the gas discharge
2. Changes in mirror spacing of optical cavity
3. Stability of cooling system
4. Competition between various rotational lines (in CO_2).

Typical specifications on amplitude fluctuations of Ar, Nd:YAG, and CO_2 lasers are ~5 percent ripple for the spectral range from ~1 Hz to 100 kHz. Two important factors for minimizing plasma instabilities are proper design of electrode geometry and ballast impedance. Fractional changes in cavity resonance frequency f and mirror spacing L are equal, i.e., $\Delta f/f = \Delta L/L$. This expression relates the frequency stability of the laser to temperature fluctuations in the cavity for known thermal expansion coefficients of the mirror spacer material.

Linewidths (Doppler in case of gas lasers), f_1, and cavity mode spacing, c/2L, for the lasers of Table 23 are shown in Table 27:

Table 27

LINEWIDTHS AND CAVITY MODE SPACING

Laser	L (cm)	c/2L (MHz)	f_1 (MHz)
Argon	50	300	3500
Nd:YAG	10	1500	~10^4 (1.06μ)
CO_2	100	150	60

Whenever f_1 is less than c/2L, large output fluctuations can occur, as the cavity resonances sweep across the laser (Doppler) linewidth, because of acoustic or thermal disturbances. For this reason, a CO_2 cavity somewhat longer than a meter may be desired to increase stability of TEM_{00} oscillations; a folded cavity ~2 or 3 meters effective length should be sufficient. The CO_2 laser oscillation must be restricted to a single rotational transition. Without such control, the spectral content of the CO_2 output (i.e., with respect to the various rotational lines) will be time dependent.

Long-term alignment of the optical cavity in a spaceborne laser is essential. Means of periodically checking and adjusting the cavity mirror alignment are recommended.

1.5 Laser for Pulsed Optical Beacon

Analysis of the optical beacon system (Chapter 4, Section 4) shows that a Q-switched, high peak power laser may be the preferred transmitter type. The desired pulse repetition rate must be ≳100 p/s if the beacon servo system is to compensate for system vibrations up to 10 Hz. The proposed laser angular divergence is ~10^{-4} radian and the average output power ≳1 to 10 W.

Q-switched lasers available for the beacon application are the 0.69 micron ruby, 1.06 micron Nd:glass, and 1.06

micron pulsed Nd:YAG. Typical characteristics of all three lasers are 10 to 20 ns pulse width, beamwidth 1 to 5 mrad ($\lesssim$1 cm beam diameter), and maximum average power 1 to 10 W. The maximum pulse repetition rate of ruby and Nd:glass, however, is 10 to 20 p/s, which appears to be a basic limitation imposed by relaxation phenomena in the laser crystal.* Nd:YAG, on the other hand, has been pulsed up to 50 p/s at 20 ns pulse width and 40 m joule/pulse (2 W average power).[26] An Xe flash lamp was used; lamp life was not studied in detail but $\gtrsim$50 to 100 hours of life have been observed in preliminary tests. Beam divergence as low as 10^{-4} rad can be obtained with an appropriate telescope. The basic limitation on repetition rate in Nd:YAG is crystal cooling; flow of a liquid coolant in direct contact with the laser rod is presently most efficient. Rates up to at least 100 p/s should be achievable with the appropriate design.

Operation of the Nd:YAG with SHG at 0.53 micron would permit use of more efficient detectors and therefore higher overall performance of the beacon system. Efficiency of SHG crystals are high with Q-switched laser operation,[2] but life is limited by optical damage[15] to the SHG material. Recent results[3] with the new material $Ba_2NaNb_5O_{15}$ indicate that it would permit ~100 percent SHG Q-switched conversion efficiency without damage to the crystal by the high intensity beam.

Pulsed Nd:YAG is recommended as the best Q-switched laser presently available for a high (~100 p/s) repetition rate optical beacon. Best performance would be provided by operation in the SHG mode at 0.53 microns; such a SHG laser with $\gtrsim$1 W average output power is not known to exist but appears feasible with appropriate developmental effort.

1.6 Mode-Locked CW Lasers

Lasers which produce a pulsed output with a high repetition rate and narrow pulse width can be advantageous for pulse position modulation (Chapter 4, Section 5) or for reduction of background noise effects by gating pulse detection circuits at the receiver. The mode-locked CW laser[27] will produce such an output. Locking can be obtained by placing an internal time-varying loss, usually an ultrasonic grating, inside the cavity. The fundamental locking frequency is c/2L, which is 150 MHz for a cavity length L of 1 meter. Subharmonic lock frequencies nc/2L, where n is an integer, are achieved by appropriate adjustment of the driving frequency of the time-varying loss. Pulsewidth τ_l of the fundamental locked mode is independent of the cavity length L and is given by

$$\tau_l = 1/f_1$$

where f_1 is the line width (Doppler width in the case of gas lasers) discussed in Section 1.4. For argon and Nd:YAG lasers, f_1 is 3.5 and 15 GHz, respectively, yielding subnanosecond values for τ_l. Mode locking generally is precluded for the CO_2 laser, since f_1 = 60 MHz and usually only one longitudinal mode will oscillate for a given rotational transition.†

Ratio of the peak-to-average power in a mode-locked laser is N, where N is the number of longitudinal modes oscillating. If f_1 is the linewidth for which the net gain of the laser is greater than unity, then $N = 2Lf_1/c$.

Mode-locking in the 0.53 micron SHG Nd:YAG laser can produce a substantial enhancement in the average power output because of the square-law relation between fundamental and second harmonic power levels. Smith[28] has shown that mode-locking increases the average 0.53 micron SHG output power by $(2N^2 + 1)/3(2N-1)$, which approaches N/3 for large N. Ratio of the peak-to-average SHG mode-locked power is 3N/2 for $N \gg 1$[28]. If the 1.06 micron linewidth f_1 is 15 GHz and the cavity length L is 30 cm, the number of longitudinal modes N is 30.

2. OPTICAL MODULATORS

In Chapter 4, Section 5, it is concluded that the optimum methods of modulation for wide-band, deep-space optical communication are biorthogonal phase shift for coherent detection and binary polarization for direct detection. To achieve a modulation rate > 1 Mb/s, an objective in this study, nonmechanical modulation techniques must be employed, e.g., electro-optic, acousto-optic, or magneto-optic interactions. Modulator driver power and/or insertion loss at present is lowest, by a wide margin, for reactive (optically nonabsorbing) electro-optic and acousto-optic modulators.‡ The following discussion therefore will be restricted to these two types. At present, electro-optic modulators provide superior performance for the application under consideration. Modulator devices will be evaluated primarily in terms of digital modulation, drive power, bandwidth, insertion loss, and optical wavelength. Size and weight will be considered only qualitatively in view of the early stages of relevant device development.

*Special barium-borate glass with Nd has permitted low average power operation up to 200 p/s with 1 m joule estimated energy per pulse at 30 p/s. See Kamogawa et al, Japan J. Appl Phys, 5 (1966), p 449.

†Experimental demonstration of a mode-locked CO_2 laser was described by E. Caddes and others at the Electron Device Meeting, Washington, D.C., in October 1967, but efficiency and peak power limitations appear to require further study.

‡One possible exception is a YIG (yttrium-iron garnet) magneto-optic modulator described by R.C. LeCraw of BTL at Intermag Conference, Stuttgart, Germany in April 1966. Further development is required to establish feasibility for deep space system use. In the present form, this modulator is promising only in the near infrared spectral region between 1.2 and 1.3 microns.

2.1 Reactive Electro-Optic Modulator Materials

The electro-optic theory of crystals, characteristics of available materials, and modulator design are treated extensively in the literature and will be discussed only briefly here. Reference 29 gives a recent review of the field.

A parameter basic to modulator design is the change of refractor index, Δn, with applied electric field E. Both Δn and E are vector quantities. Directions of E and the incident light polarization, with respect to crystal axes, consequently must be carefully selected for optimum modulation. For the linear electro-optic effect, where quadratic and higher order terms are neglected, Δn is given to a good approximation[29] by the electro-optic tensor

$$\Delta \frac{1}{n_i^2} = r_{ij} E_j \qquad (1)$$

where r_{ij} is the linear electro-optic coefficient. The indices i and j can run from 1 to 6 and 1 to 3 respectively; however, many r_{ij} values are zero depending on crystal class (or point-group symmetry). Equation (1) must be summed over ij, but in many cases of practical interest all but one can be neglected. For this case $\Delta n \sim -n^3 rE/2$. The optical phase change $\Delta \phi$ produced by the applied electric field is

$$\Delta\varphi = \frac{2\pi L \Delta n}{\lambda} \simeq -\frac{\pi L n^3 r E}{\lambda} \qquad (2)$$

where L is modulator crystal length in the direction of light propagation and λ is the vacuum wavelength. Equation (2) applies directly to the case of phase modulation where the incident polarization is parallel to either the ordinary or extraordinary axis (Section 2.3). With polarization modulation, the incident beam is typically polarized at 45 degrees to the two axes (Section 2.2) and Equation (2) becomes

$$\Delta\varphi = -\frac{\pi L E}{\lambda} \cdot \left(n_o^3 r_o - n_e^3 r_e\right) \qquad (3)$$

where o and e refer to the ordinary and extraordinary polarization directions. The field E = V/D, where V is the applied voltage and the cross section of the crystal, is a square of dimension D on a side. Another important design parameter is the half-wave voltage V_π required to produce a phase change of π for a crystal in which L = D, i.e., a cube-shaped modulator. V_π is basic since it is independent of D; the relations for V_π can be obtained from Equations (2) and (3) when $\Delta\phi = \pi$ and L = D which yield:

$$V_\pi = \frac{\lambda}{n^3 r} \quad \text{(phase modulation)} \qquad (4)$$

$$V_\pi = \frac{\lambda}{n_e^3 r_e - n_o^3 r_o} \quad \text{(polarization modulation)} \qquad (5)$$

For the $LiTaO_3$ and $Ba_2NaNb_5O_{15}$ crystals discussed later in Section 2.2, $r_e = r_{33}$, $r_o = r_{13}(LiTaO_3)$, and $r_o = r_{23}$ $(Ba_2NaNb_5O_{15})$. Note that for a crystal of length L, the half-wave drive voltage $V_{D\pi}$ is

$$V_{D\pi} = \frac{D}{L} V_\pi \qquad (6)$$

A diagram of a reactive electro-optic modulator is shown in Figure 76. Crystal cross section is assumed to be square, with a dimension D on each side. Matching of the laser beam to the modulator is provided by a lens. Optimum coupling is achieved, in the case of a Gaussian beam cross section, when the beam geometry inside the crystal is that of a confocal resonator of length L. For this condition the ratio D^2/L of the crystal dimensions is given by[29]

$$D^2/L = 4\lambda/n\pi \qquad (7)$$

In practice D^2/L should be somewhat larger, by approximately a factor of 10, than $4\lambda/n\pi$ to allow margin for effects such as aberrations and mechanical tolerances.[29]

Input modulation signal V(t) is applied by a driving circuit to the crystal electrodes. The drive can be either single-ended, as indicated in Figure 76, or push-pull. The choice depends on circuit and mechanical design considerations. Driver circuit output power is basically all reactive because of the high resistivity of electro-optic materials of interest. Reactive driver power, P_D, for an applied field E = V/D and bandwidth B is $CV^2 B/4$, where C is total drive capacitance. If the dielectric constant is ϵ, ϵ_o is permittivity of free space, and fringing fields are neglected, then $C = \epsilon_o \epsilon L$ and P_D becomes for $V = V_{D\pi}$ using Equation (6)

$$P_D = \frac{CBD^2 V_\pi^2}{4L} = \frac{\epsilon_o \epsilon B D^2 V_\pi^2}{4L} \qquad (8)$$

At optimum coupling $D^2/L \propto \lambda/n$ (Equation 7), from which it follows that, for a given phase shift π and a crystal with square cross section, dependence of drive power on bandwidth, optical wavelength, and materials constants is, for phase modulation,

$$P_D \propto \left(\frac{D^2}{L}\right)\left(B\epsilon V_\pi^2\right) \propto B\epsilon\lambda^3/n^7 r^2 \qquad (9)$$

and, for polarization modulation,

$$P_D \propto \left(\frac{D^2}{L}\right) B\epsilon V_\pi^2 \propto B\epsilon\lambda^3/n \left(n_o^3 r_o - n_e^3 r_e\right)^2 \qquad (10)$$

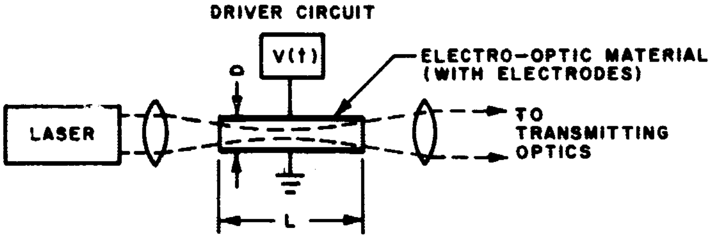


Figure 76. Reactive elec..o-optic modulator

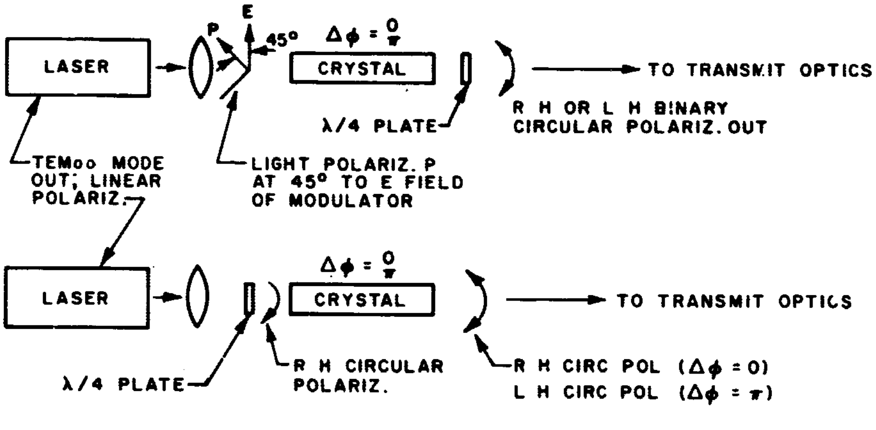


Figure 77a. Schematic diagram of binary polarization el:ctro-optic modulator

It should be noted that Equations (9) and (10) are for ideal optical coupling; in practice P_D must be higher, perhaps by a factor of 10, to allow for margins. Note also the advantage of shorter wavelengths ($P_D \propto \lambda^3$).

Material properties of greatest importance for electro-optic modulator applications are:

1. Dielectric constant ϵ, which affects driver capacitance and propagation velocity of electric fields through the crystal
2. Refractive index n, which is inversely proportional to light-beam transit time through the modulator; as noted later, transit-time effects can be neglected for bandwidths ~$10^6 s^{-1}$
3. Electro-optic coefficient r, which, together with n, determines the electrically induced change in refractive index Δ n (see Equation 2)
4. Optical and electrical losses, as given respectively by the absorption coefficient α and the dielectric loss tangent δ
5. Physical hardness, freedom from growth defects, capability of single crystal growth in large boule sizes, and minimum hygroscopicity, all of which are necessary for fabricating durable and long-life modulator crystals of usable size
6. Resistance to optically induced damage[30] by intense focused laser beams
7. Thermal expansion and thermal conductivity coefficients which affect thermally induced birefringence due to thermal changes in optical length or thermal gradients.

Requirement 5 eliminates many of the known electro-optic materials as candidates for a modulator. Of those that remain, the highest overall performance materials for the 0.5, 0.69, 1.06, and 10.6 micron wavelengths of particular interest in this study are, at present, $LiTaO_3$ and $Ba_2NaNb_5O_{15}$ for the 0.5 to 1.06 micron range and GaAs at 10.6 microns. It should be noted, however, that electro-optic materials research is very active and that availability of even higher figure of merit crystals is probable in the not too distant future. $Ba_2NaNb_5O_{15}$ is a new material[31] that promises somewhat higher performance than $LiTaO_3$. Relative comparison of $LiTaO_3$ to KDP, KTN, and $LiNbO_3$ appears elsewhere[32] and serves as a basis for material selection here for the binary polarization modulator. Parameters, including ϵ, n, and r for T ~ 300°K are given in Table 28 for the three selected materials. Precision of ϵ and r data in the table is one or at best two significant figures because of variations[29] in measurement technique, temperature and wavelength. However, Table 28 will be sufficient for modulator evaluation within the wavelength ranges given and for bit rates $\lesssim 10^6$ to $10^7 s^{-1}$. Optical absorption is typically a few decibels for the materials listed in Table 28 but varies appreciably between samples, with optical coupling, and with the quality of antireflection coatings on the crystal ends. Losses of 1.5 dB for a 1 cm long $LiTaO_3$ crystal[32] at 0.63 micron and <0.1 cm^{-1} for GaAs[33] at 10.6 microns have been observed experimentally. The ratio D^2/L in the $LiTaO_3$ sample for which the 1.5 dB insertion loss was obtained was about a factor of 10 larger than the theoretical limit $4\lambda/n\pi$ (Equation 7), which suggests that a tradeoff may exist in practice between drive power (Equation 8) and optical loss. Values reported for loss tan δ of GaAs are <0.01 when measured at 9.3 GHz[29] and 0.001 at 2.5, 5.6, and 10 GHz.[33]

2.2 Binary Polarization Electro-optic Modulator (0.5 through 1.06μ)

Binary polarization modulation with direct (incoherent) detection is proposed for 0.5 or 1.06 micron wavelengths (Chapter 4, Section 5). Most promising modulator performance is currently offered with a $LiTaO_3$ or $Ba_2NaNb_5O_{15}$ electro-optic material in the configuration shown in Figure 76. The E field is applied along the c-axis of the crystal, which must induce a phase change of π (given by Equation 3) to change from a binary 0 to a 1. Incident polarization can either be linear, at an angle of 45 degrees to E or c-axis (Reference 32 is an example) or circular. Circular polarization is desired at the modulator output to avoid transmitter-receiver angular alignment. Binary circular polarization output (RH or LH) is achieved as shown in Figure 77a. A $\lambda/4$ plate is needed for either arrangement, on the exit side for incident linear polarization 45 degrees to E and at the entrance for circular polarization at the modulator input. Voltages applied from the driver are zero for a binary 0 and $V_{D\pi}$ ($\Delta\phi = \pi$) for a binary 1. Phase shift at V = 0, due to natural birefringence, strains, etc., must be $2m\pi$ in both cases, where m is an integer. This zero field condition can be obtained by at least two methods:[32] (1) by appropriate adjustment of the absolute crystal temperature or (2) with a Babinet-Soleil compensator at the modulator output. Mechanical motion is required for adjustment of (2).

Precision control of the crystal temperature, T, is essential to maintain a constant zero-field phase shift at $2m\pi$. Calculations[32] for a binary linear polarization output from $LiTaO_3$ show that ΔT must be $\leqslant \pm 0.045°C$ if the ratio of 1 to 0 is to be $\geqslant$ 20 dB with a 1-cm long crystal. This ΔT requirement should be substantially valid also for circular polarization output. In short, T must be maintained constant to ~10^{-2}°C for the above polarization modulator.

Power requirements include those of the modulator driver and temperature control circuits. $LiTaO_3$ polarization modulators have been designed and operated up to 10^8 Hz bandwidth.[32] Typical design parameters (see, for example, Reference 32) are an aspect ratio L/D = 40 to 80, D = 0.02 to 0.025 cm, and C = 5 pf. Equation (8) and V_π values of Table 28 show that the magnitude of the

Table 28

CONSTANTS OF ELECTRO-OPTIC MATERIALS

	$LiTaO_3$*	$Ba_2NaNb_5O_{15}$†	GaAs^{c‡}
ϵ_3	40	51	12
n_e	2.180 (0.63μ)	2.256 (0.53μ) 2.175 (1.06μ)	–
r_{33}(cm/volt)	3 X 10^{-9} (0.63μ)	5.4 X 10^{-9} (0.53μ)	–
n_o	2.176 (0.63μ)	2.370 (0.53μ) 2.261 (1.06μ)	3.3
r_{13}(cm/volt)	7 X 10^{-10}(0.63μ)		–
r_{23}(cm/volt)	–	1.3 X 10^{-9} (0.53μ)	–
r_{41}(cm/volt)	–	–	1.6 X 10^{-10}
V_π (volts), Equations (4) and (5)	2800(0.63μ)	1570 (0.63μ)	1.6 X 10^5 (10.6μ)§
Range of transparency	0.4-5μ	0.4-5μ	0.9-16μ

* References 29 and 32.

† Reference 31.

‡ References 29, 33, and 34.

§ $V_\pi = (\frac{\sqrt{3}}{2})\lambda/n^3\gamma_{41}$ for GaAs 111 crystal axis considered in Section 2.3.

reactive output power of the driver for B = 10^6 Hz is P_D ~10^{-2} to 10^{-4}W. The required peak-to-peak driver voltage $V_{D\pi}$ is (D/L)V_π which at L/D = 50 and λ = 0.5 micron yields $V_{D\pi}$ ~45 and 25 V p-p for $LiTaO_3$ and $Ba_2NaNb_5O_{15}$, respectively. Low power (~1 to 10W total input) transistor circuitry can readily provide the above P_D and $V_{D\pi}$ levels. At L/D = 50, λ = 0.5 micron, D = 0.025 cm, and n = 2.2, the ratio D^2/L = 22 λ/n or 17 times the minimum value of 4λ/nπ from Equation (7), which should be more than adequate margin. Transit times are sufficiently short that no traveling wave structures are needed at B~10^6 Hz. Constant temperature control technology is well developed and $\Delta T < 10^{-2}$°C is within the state of the art. Estimated input power for an optimized temperature controller operating with crystal temperature near ambient is ~1 to 10W. The size of such a modulator, based on those built in the laboratory, is the order of 0.04 ft^3 (a cube 4 in. on a side) and weight ~5 to 10 lb. Data available on optical damage with 4880A light show that at least 500W/cm^2 (0.2W focused to a 0.2 mm spot) and 750W/cm^2 (0.3W focused to a 0.2 mm spot) can be transmitted through $LiTaO_3$ and $Ba_2NaNb_5O_{15}$ respectively with no observable damage. Care in heat treatment and poling of $LiTaO_3$ must be taken to achieve optimum damage resistance.[30,32] Care must also be taken to minimize photoconductive effects which can produce a nonuniform transverse distribution of the applied electric field.[32] Acoustic resonances, a result of the natural piezoelectric properties of electro-optic materials, are minimized by appropriate mechanical design for damping such oscillations.[32]

The binary polarization modulator can be summarized as follows:

Material	Choice
$Ba_2NaNb_5O_{15}$	1
$LiTaO_3$	2

Future improvements in crystal growth technology are needed to provide the requisite modulator efficiency and life.

Crystal aspect ratio L/D:>50(D~0.02 – 0.04 cm)

Incident polarization: linear (at 45 degrees to c-axis) or circular

Exit polarization: circular R H (0) or L H (1)

Zero field phase shift: 2mπ(m = integer)

Crystal temperature control: ~±10^{-2} °C

Reactive driver output power: ~10^{-2} – 10^{-4} W

Driver output voltage (D/L) $V_\pi = D\lambda/L(n_e^3 r_e - n_o^3 r_o)$: ~25 volts ($Ba_2NaNb_5O_{15}$ for λ = 0.5 micron and L/D = 50)

Power input to driver: ~1to 10W

Power input to temperature controller: ~1 to 10W

Size: $\lesssim$ 0.04 ft^3

Weight: ~5 to 10 lb

Electrode structure: bulk modulator OK at B ~ 10^6 Hz (traveling wave structure not needed)

Insertion loss: < 1 dB (with antireflection coatings on crystal).

2.3 Biorthogonal Phase Electro-Optic Modulator (10.6μ)

Materials that are transparent at 10.6 microns and have a high known electro-optic coefficient are GaAs,[29,33,34] Se,[35] CdS,[29] and ZnTe.[29] The figure of merit n^3r (Equations 2 and 4) is highest for Se; n^3r values for GaAs and CdS are comparable. The crystal quality and/or size of Se[35], CdS, and ZnTe is at present inadequate for practical modulators. Consequently, of the four materials mentioned above, GaAs, doped with Fe or Cr to increase the resistivity to reduce dissipation losses, currently yields the best performance at 10.6 microns. It is GaAs that will be considered in detail in this section as an electro-optic phase modulator.

Figure 77b is a schematic diagram of the binary phase shift modulator. Incident polarization to the crystal is linear, which is changed to circular polarization at the modulator output. The structure of GaAs is cubic zincblende. Maximum phase shift is attained when the E-field and light polarization are parallel and along a 111 crystal axis.[29] The relevant electro-optic coefficient is r_{41} (Table 28). Modulator crystal is a parallelepiped with the following orientation:[29]

1. Light propagates normal to a 111 plane
2. Light polarization is along 111 axis and parallel to E
3. Binary E-field pulses change phase of the coherent light by zero (binary 0) or by π (binary 1)
4. Linearly polarized wave exiting from modulator into circular polarization is converted by a quarter wave plate, e.g., CdS 5λ/4 plate (Figure 77b).

Drive power P_D varies as λ^3 near optimum coupling (Equation 9); hence there is an important tradeoff between P_D and λ. If a factor of 10 margin is allowed in the D^2/L ratio of Equation (7), i.e., $D^2/L = 40\ \lambda/n\pi$, then L ~10 cm for a crystal thickness D of 2 mm. Feasibility of values of D much smaller than 2 mm is doubtful because of optical absorption, aberrations, and minimum allowable alignment tolerances. Total capacitance of such a modulator is typically 15 pf. From Equations (6) and (8), the reactive driver power at 10^6 Hz bandwidth for a π phase shift* then is 40 watts and the required drive voltage $V_{D\pi} = 3.2$ kv. Although this is a high voltage and represents difficult circuit design problems, the dissipative power can be quite low. Resistivities as high as $10^8\ \Omega$ cm have been obtained in Cr-doped GaAs crystals for which the optical quality and transmission were comparable to or better than for Fe doping. Dissipative losses in a 0.2 cm × 0.2 cm × 10 cm modulator with 3.2 kV across the electrodes would be ~4×10^{-4} watts.

Therefore, the critical problems in a 10.6 micron GaAs modulator are:

1. Pulse circuit design that provides ~4 kV peak voltage into a reactive load
2. Prevention of electrical breakdown either through the gas surrounding the modulator or defects in the modulator crystal
3. Minimize driving circuit power requirements, e.g., with pulse circuits resonant at the pulse repetition rate.

Another consideration is variation in static birefringence due to temperature changes. This effect is much less in cubic GaAs, where $n_x \simeq n_y$, than for the tantalates or niobates of Section 2.2. Although insufficient data are available for quantitative evaluation, the permissible temperature fluctuation is expected to be one or more orders of magnitude higher than the ~10^{-2}°C required for polarization modulators in Section 2.2. Size, weight, and input power will depend strongly on the specific driving circuit design; it seems clear, however, that they will be higher (by at least a factor of two) than those of the polarization modulators discussed in Section 2.2. Further work on driving circuits and/or new modulator materials is needed for quantitative determination of volume, weight, and power requirements of an electro-optic 10.6 micron modulator.

2.4 Acousto-Optic Modulators

Interaction of light with propagating acoustic waves is reactive and therefore potentially can provide efficient wideband modulation of light. Analyses of acousto-optic phenomena, in particular the dependence of modulation index on materials constants, input power, acoustic frequency, and bandwidth, appear in recent review articles (see, for example, References 36, 37, and 38). Most of the work to date has been concerned with theory and basic materials properties rather than development of specific devices. A primary reason for this is that, until the advent of the laser, application of acousto-optic modulators was limited by the lack of intense monochromatic light sources. In addition, important advances in materials and acoustic transducer technology have been made in the past few years (see, for example, References 39 and 40). As a result, intensive acousto-optical device development is in the early stages and significant progress relative to the present state of the art can be expected in the next 5 to 10 years.

Efficient amplitude modulation is provided by the Bragg angle acoustic modulator illustrated in Figure 78a. Because

*For E and light polarization along the 111 GaAs axis, $V\pi = (\sqrt{3}/2\ \lambda/n^3 r_{41}$ (Reference 29).

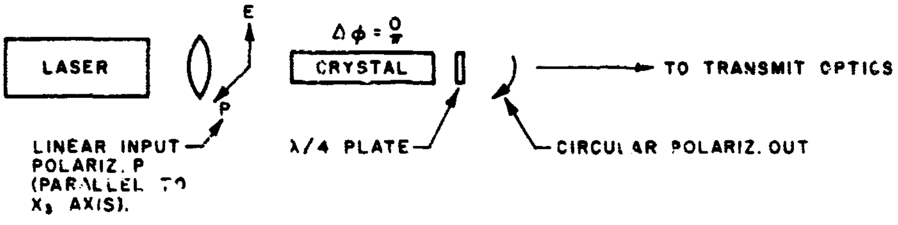


Figure 77b. Schematic diagram of binary phase shift electro-optic modulators

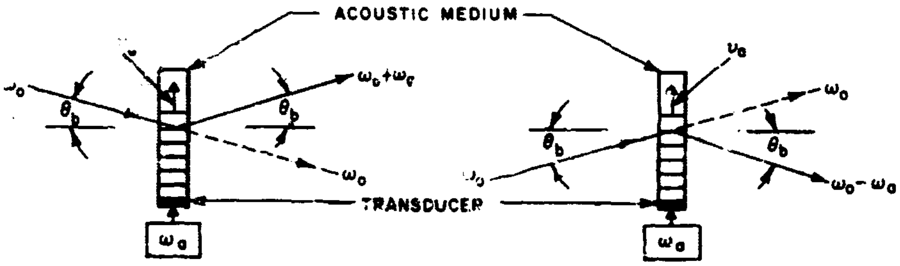


Figure 78a. Diagram showing possible light input and output configurations for Bragg acousto-optic modulators

of its many potential applications, emphasis of work to date in the acousto-optic field has been on the Bragg type. Laser light is incident at the Bragg angle $\theta_b = \lambda_o/2\lambda_a$, where λ_o is the optical wavelength in vacuum and λ_a the acoustic wavelength in the medium. Frequency ω_o of the incident light is shifted to $\omega_o - \omega_a$ or $\omega_o + \omega_a$ upon Bragg diffraction, depending on whether the transverse momentum component of the light is parallel or antiparallel respectively to the acoustic wave velocity (Figure 78a). With sufficient acoustic power P_a, all light leaving the transducer will appear in the diffracted component.

Binary polarization modulation, preferred for deep space communication at 0.5 or 1.06 microns (Section 2.2), can be achieved by utilizing the acoustically induced birefringence effect. Polarization of the Bragg-diffracted light will be at right angles to the incident light polarization for all nonisotropic noncubic crystals in the case of a transverse acoustic wave[41]. Optimum materials at present are GaP or As_2S_3 at 1.06 microns and $LiNbO_3$ or TiO_2 (rutile) at 0.5 micron; data for relevant materials parameters[39] are given in Tables 29 and 30. A possible mode of operation is binary 1 for the diffracted light and binary 0 for the direct transmitted light (indicated by dashed lines in Figure 78a).

Table 29

ELASTO-OPTIC CONSTANTS, p, OF ACOUSTO-OPTIC MATERIALS[39]

	λo (Microns)	P_{11}	P_{12}	P_{44}	P_{31}
GaP	0.63	−0.151	−0.082	−0.074	-
As_2S_3	1.15	0.308	0.299	-	-
$LiNbO_3$	0.63	0.036	0.072	-	0.178
TiO_2	0.63	0.011	0.172	-	0.0965
GaAs	1.15	−0.165	−0.140	−0.072	-
Te	10.6	0.155	0.130	-	-

Recombination of the two beams into collinear waves for propagation through the transmitting optics can be accomplished with a Koster's prism[37] as shown in Figure 78b. The following factors are of concern in this type of polarization modulator:

1. A 3 dB loss appears at the Koster prism due to the beamsplitter (Figure 78b).
2. The diffraction angle, θ_d in Figure 78b, must be maintained constant consistent with the desired pointing precision.
3. Driver power is required for ~10^6 b/s information rate.

In view of the 3-dB loss at the prism, systems performance would be comparable without the prism and if pulse modulation were used instead with the transmitted beam only. The relation between incremental changes in diffraction angle θ_d and acoustic frequency f_a is $\Delta f_a = (v_a \Delta\theta_d)/\lambda_o$; for $v_a = 4 \times 10^5$ cm/sec, $\lambda_o = 0.5$ micron, and $\Delta\theta_d = 10^{-6}$ rad (0.2 arc-sec), and the required acoustic driver stability is $\Delta f_a = 8$ kHz. For ~10^6 b/s information rate, typical f_a values are $10^7 - 10^8$ Hz, in which case $\Delta f_a/f_a$ ~10^{-3} to 10^{-4}. Acoustic power P_a in the crystal required for diffracting 100 percent of the light can be estimated from the P_a/f_a values given in Table 30, where $P_a/f_a = \lambda_0^3 \rho v^2/3.6n^7 p^2$ (see Tables 29 and 30). It was assumed during computation of these values[39] that the acoustic beam height was near its minimum value of v_a/f_s, where f_s is the information rate, that the acoustic and optical losses are negligible, and that the appropriate optimum p coefficient is used for either transverse or longitudinal wave propagation. Also, it should be noted that $P_a/f_a \alpha \lambda_0^3$, similar to Equation (10) for the electro-optic modulator. Generally f_a should be 10 to 100 times f_s, which yields values for P_a of ~5 to 60 mW at 0.5 micron ($LiNbO_3$ or TiO_2) and 1.5 to 100 mW at 1.06 microns (GaP or As_2S_3). Typical coupling losses at the transducer at present are

Table 30

CONSTANTS OF ACOUSTO-OPTIC MATERIALS[39]

Material	λo (Microns)	n	$\left(\frac{\rho}{\frac{gm}{cm^3}}\right)$	$\left(\frac{v_a}{10^5 \frac{cm}{sec}}\right)$	$\frac{n^6p^2}{\rho v_a^3}$ (X10^{-11} MKS)	P_a/f_a (mw/MHz)	Acoustical Wave Polarization
GaP	0.63	3.31	4.13	4.13	24.1	0.21	Trans.
As_2S_3	1.15	2.46	3.20	2.6	347	0.179	Long.
$LiNbO_3$	0.63	2.20	4.7	6.57	6.99	0.69	Long.
TiO_2	0.63	2.58	4.6	7.86	3.93	0.87	Long.
GaAs	1.15	3.37	5.34	3.32	46.3	0.86	Trans.
Te	10.6	4.8	6.24	2.2	4400	7.14	Long.

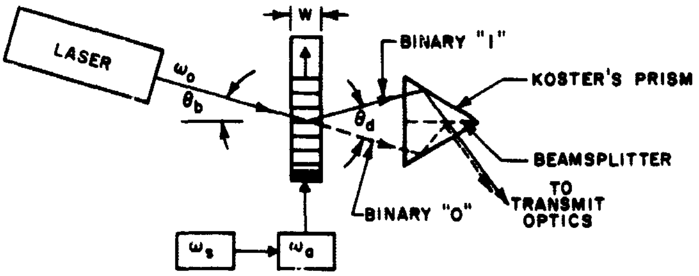


Figure 78b. Diagram showing possible light input and output configurations for binary polarization acousto-optic modulators

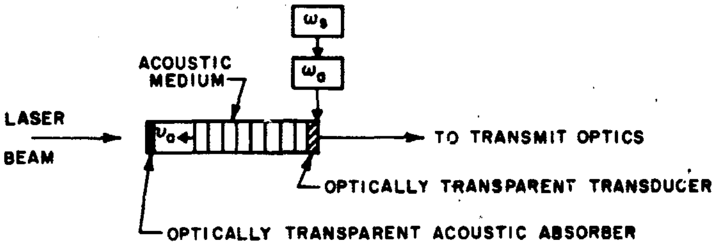

ω_s
ω_a
ACOUSTIC MEDIUM
LASER BEAM
v_a
TO TRANSMIT OPTICS
OPTICALLY TRANSPARENT TRANSDUCER
OPTICALLY TRANSPARENT ACOUSTIC ABSORBER

Figure 79. Collinear binary polarization acousto-optic modulator (Reference 39)

13 – 20 dB; if 13 dB is used, allowing for future advances, then total acoustic driver output power P_D is:

λ(microns)	$f_a = 10^7$ Hz	$f_a = 10^8$ Hz
0.5	$P_D = 0.1$ W	$P_D = 1$ W
1.06	$P_D = 0.03$ W (As_2S_3)	$P_D = 0.3$ W (As_2S_3)
	~0.2 W (GaP)	~2 W (GaP)

GaP would provide better performance at 0.53 micron than $LiNbO_3$ or TiO_2, but requires cooling to ~77°K because of band-edge absorption.[39] Such cooling could represent a substantial penalty in deep space applications.

A second, and more speculative, mode of acousto-optic binary polarization modulation is collinear transmission[39] of the optical wave parallel or antiparallel to the acoustic wave as illustrated in Figure 79. This mode has the advantage that no Bragg deflection occurs. Therefore it would not introduce the 3-dB beamsplitter loss or possible pointing errors due to drifts in ω_a, relative to the Bragg type of Figure 78b. Three practical problems of the collinear modulator, however, are (1) fabrication of an efficient optically transparent transducer, (2) fabrication of an optically transparent acoustic absorber, and (3) transit time of the acoustic wave which must be $\leqslant$ 1 bit period. For a velocity $v_a \sim 4 \times 10^5$ cm/sec, the maximum length of the acoustic medium is 0.4 cm at 10^6 b/s. The optically transparent coatings are probably not available with present art.

The preferred modulation at 10.6 microns in a deep space link is biorthogonal phase shift (Chapter 4, Section 5). This can be attained by the acousto-optically induced change Δn in the refractive index n as shown in Figure 80. The relation[36 to 38] for Δn is

$$\Delta n = \frac{n^3 ps}{2} \qquad (11)$$

where p is the elasto-optic constant and s is the strain. Acoustic power P_a is defined[37] as

$$P_a = \frac{1}{2} \rho \, v_a^3 s^2 wh \qquad (12)$$

where ρ is mass density of the medium, w the width of the acoustic beam at its waist, and h the beam height. Phase change ΔΦ, given by Equation (2), is ΔΦ = 2πwΔn/λo. With biorthogonal phase shift modulation, ΔΦ = π or 0, thus (Δn) max = λo/2w which, when combined with Equations (11) and (12), yields

$$P_a = \frac{\rho \, v_a^3}{n^6 p^2} \; \frac{\lambda_o^2 h}{2w} \qquad (13)$$

Best materials at 10.6 microns currently[39] are GaAs and Te, for which the pertinent constants and the figure of merit $n^6 p^2/\rho v_a^3$ are given in Tables 29 and 30. Material constant data on GaAs in the two tables were measured at 1.15 microns; it will be assumed that the values hold also at 10.6 microns in the subsequent discussion of this section. A requirement on this type of phase modulator is that the optical beam diameter D_o in the interaction region must be appreciably less than, say, ~0.2, the acoustic wavelength λ_a. Values of λ_a for $f_s = 10^6$ bits/sec are (from Table 30) 0.3 cm and 0.2 cm respectively for GaAs and Te. If $D_o = 0.2\lambda_a$, then the desired respective optical spot sizes are 0.6 mm and 0.4 mm, which are achievable with 40 to 60 cm focal length lenses for a 1-cm diameter diffraction-limited 10.6 micron laser beam. The respective acoustic powers given by Equation (13) for h = w and $\lambda_o = 10.6$ microns are $P_a \simeq 10$ W and 1000 W for Te and GaAs respectively. Acoustic losses in Te of a few dB/cm have been reported[42] at 300 MHz acoustic frequency; however, only very short acoustic columns are required (Figure 80) since $D_o << \lambda_a$; thus acoustic losses should not be a severe problem. Also, the acoustic width w can in principle be small (0.2 – 0.4 λ_a) which would keep absorption losses << 1 dB ($\alpha \sim 0.1$ cm^{-1} for GaAs and < 0.5 cm^{-1} for Te).

Te is the best material for the biorthogonal phase modulator at 10.6 microns. Acoustic power P_a is less, by $\sim 10^{-2}$, compared to the next best material, GaAs. If the transducer loss is 13 dB, the driver power of a Te acoustic-optic modulator would be ~ 200 watts. Focusing of the 10.6 micron CO_2 laser beam to ~ 0.4 mm diameter in the interaction region is required. Work may be necessary to reduce absorption losses in Te at 10.6 microns before such a modulator would be practical.

3. TRANSMITTING OPTICS

The optics aboard the vehicle must satisfy demanding and, to some extent, contradictory requirements. The surfaces must be figured and maintained to diffraction-limited tolerances (~λ/50 at the wavelength used) to realize the full capability of antenna gains at optical frequencies. This must be done under remote and hostile conditions. Although the distortions produced by gravity are relaxed, the optics are exposed to meteoroid damage and to a harsh thermal environment. The removal of convective cooling means that heat is exchanged by conduction (where glassy materials conduct poorly) and by radiation (where reflecting surfaces radiate poorly). Since visible wavelengths are the smallest of interest, tolerances are most demanding in the visible range and are therefore considered here.

It is also important to estimate weight and costs of the various aperture sizes of transmitters that might be put on the vehicle. The difficulty here is that there is relatively little hard experience to draw on, and costs especially are uncertain.

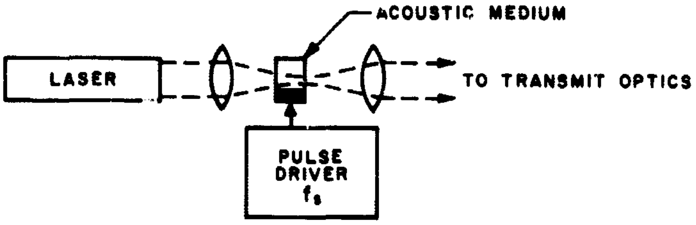


Figure 80. Biorthogonal p.ia.ic modulaior using acousto-optic medium

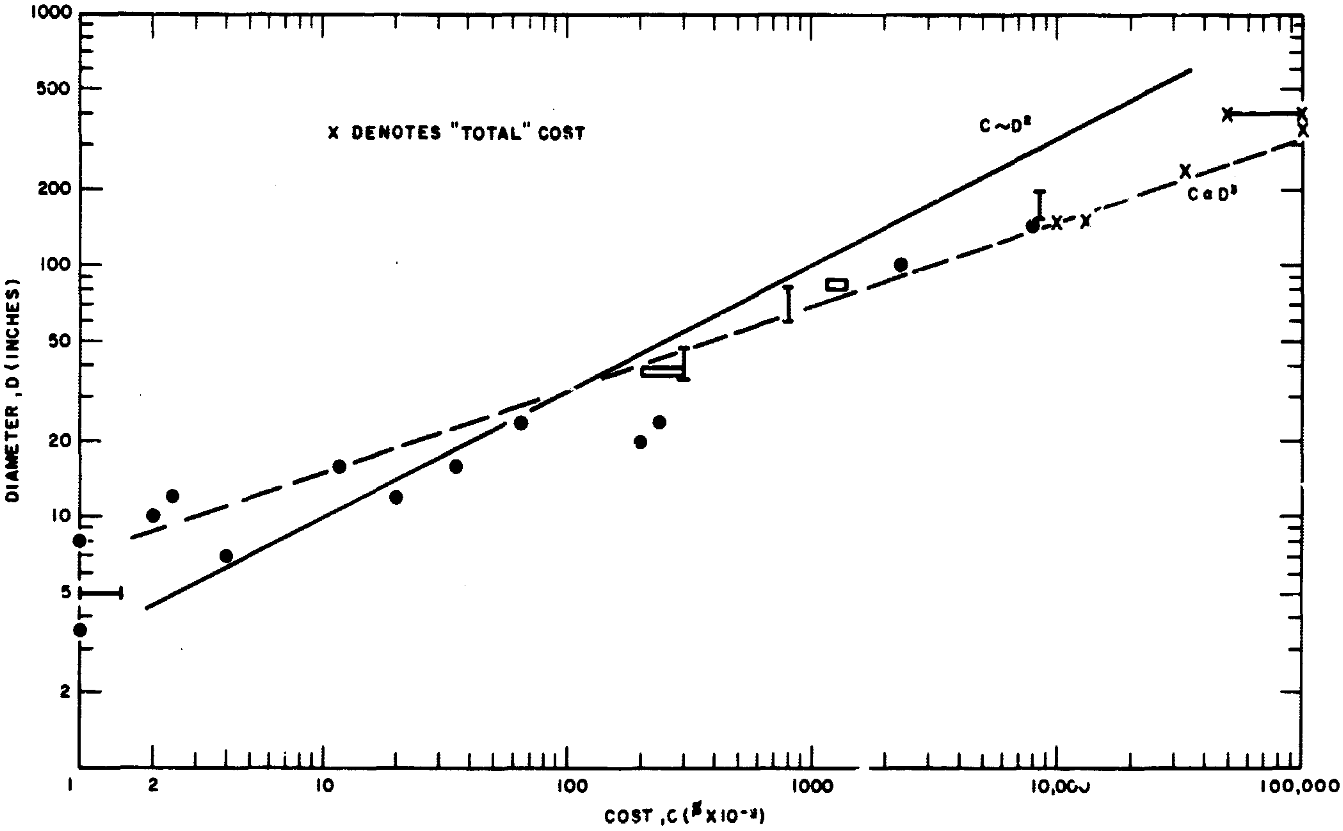


Figure 81. Telescope cost as a function of diameter

3.1 Transmitting Optics – Costs

Cost, of course, very much depends on the degree of quality which will be required, and this depends especially on the wavelength used. Nevertheless, experience with Earth-based telescopes has shown that cost is dominated by aperture size. Moreover, the quality of optics required is, at least below a certain size, similar to that of Earth telescopes: the optical system must be diffraction-limited. (And, as shown below, those large sizes which are not made diffraction-limited do not deviate significantly from the cost dependence of smaller, diffraction-limited telescopes.) Thus, it may be reasonable to estimate the cost of the transmitting optics from the costs of astronomical-quality Earth telescopes. Also, the mounting problems are different, but perhaps not unlike in difficulty. Although the space unit is gravity-free in operation, it must survive a rocket launch. The Earth telescope costs include the mounting and, for all but the small sizes, the mounting involves considerable sophistication in order to minimize gravity distortions.

In 1964, the Whitford Committee summarized[43] the construction costs of existing astronomical telescopes to provide perspective to their recommendations for future facilities. The plotted data indicated that cost was proportional to collecting area. However, barely more than one decade of diameter was covered, the projections for future giant telescopes were not included, and at least one point was in error. Therefore, current prices and recent estimates for two decades of diameter are shown in Table 31 and plotted in Figure 81. Thus, for example, the 200-inch Hale telescope is omitted in favor of recent estimates for large (>100-inch) telescopes.

Such a conglomerate necessarily includes telescopes of substantially different characteristics, and these differences should be understood before Figure 81 is interpreted. A variation in the quality with increasing size is incorporated in all such studies; good small telescopes are diffraction-limited, good large ones are not (for visible light). The change generally occurs at about 10 or 20 inches. The large ones are simply made good enough that atmospheric seeing limits resolution. The points corresponding to "low" quality small telescopes (surfaces good to perhaps $\lambda/4$) are included to indicate the functional dependence of cost if the surface quality of the large telescopes is approximated. Another difference is that the total cost of large telescopes usually includes non-telescope items ranging from land to ancillary instruments. Where these could not be separated, total cost values are denoted by a cross in Figure 81. Bars and boxes correspond to value ranges.

It is clear that cost varies with diameter more rapidly than by the square. Allowance for the differences between telescopes in no way allows a square law to fit well. If points beyond the range (16 to 200 inches) of the Whitford study[43] are ignored, a second power fit seems less bad, but still not accurate. An overall fit shows that the exponent should be at least 2.5. It is interesting that the large telescope "total" values fit the same straight line well. If they were corrected to telescope-only values, then a square dependence might fit the large-telescope end. However, then a single law could not hold for all sizes because of the constraint set by the small telescopes. Indeed, if the values for "low quality" small telescopes are used to estimate the dependence for approximately constant surface quality, then the cost seems to vary with about the cube of diameter.

These results are in contrast with the rather uniform opinion of astronomers that the square dependence fits well. Rule, at the 1965 Tucson Symposium[47] on large telescopes, observed that the simple square dependence had been used throughout the symposium, and he contrasted this with radio telescopes where the exponent is 2.5 to 2.7. It would appear that the dependence of cost on diameter for optical telescopes actually is not very different from that for radio telescopes.

3.2 Transmitting Optics – Weight

The weight of a telescope to be put in space very much depends upon the nature of the mission. The Orbiting Astronomical Observatories (OAO's) had some relatively light telescopes[48] (8 inches, 28 pounds; 16 inches, 74 pounds), but these were not of diffraction-limited quality in the visible and were not complete communications units. Perkin-Elmer has made individual designs[49] for a number of aperture sizes of diffraction-limited, complete communications telescopes. This appears to be the most accurate and relevant set of weights available, and they are listed in Table 32. No simple power law relation between diameter and weight will fit accurately the data in the table. The smaller sizes would employ single-element primary mirrors, but the 80 and 120 inch mirrors would involve active optics; they would be segmented, monitored, and adjusted. The feasibility of this technique has been demonstrated by Perkin-Elmer.

The question of how large a primary mirror is wanted in a deep space vehicle is only one parameter in the necessary trade-off study, but there is much to recommend a diameter of about one meter. This is about the size that can be accomplished before the weight and complexity of active optics are required. The size of the vehicles that would be used first could contain 1-meter optics gracefully. To make a meaningful improvement in telescope collecting area, or gain, would involve, say, doubling the diameter. Scaling the rest of the telescope up to a 2-meter primary leads to a structure so massive that a very substantial vehicle would be required. It is noteworthy that Stratoscope II, a balloon-flown telescope, had a 36-inch primary which was successfully figured to $\lambda/50$. Thus, the 1-meter size represents well-established technology.

Table 31
TELESCOPE COSTS

Telescope Aperture (inches)	Cost ($\$10^3$)	Location or Source	Quality	Comments
3.5	1	Questar	High	Large ratio of cost to quality
7	4	Questar	High	Large ratio of cost to quality
5	1-1.5	Tinsley	Moderate	
8	1	Tinsley	Low	
12	2.4	Tinsley	Low	
12	20	Fecker Instruments	High	
10	2	Celestron		
16	11.5	Celestron		
16	35	Boller and Chivens	Observatory quality	
24	65	Boller and Chivens	Observatory quality	
36-40	200-300	Boller and Chivens	Observatory quality	
80-90	1, 150-1, 400	Boller and Chivens	Observatory quality	
20	200	Baker-Nunn		Satellite tracking
24	240	Grubb-Parsons		Satellite tracking
36-48	300	Whitford study[43]		Basic telescope*
60-84	800	Whitford study[43]		Basic telescope*
98	2,800	Greenwich Observatory, England		Basic telescope
150-200	8,500	Whitford study[43]		Basic telescope (about 18,500 total cost)
144	8,000	European southern observatory[44]		Basic telescope estimate by Professor Heckmann
150	10,000	Cerro Tololo,[45] Chile		Probably total cost*
150	13,000	New South Wales, Australia[45]		Probably total cost*
240	33,000	Zelenchut,[46] Russia		Probably total cost
350-400	About 100,000	Whitford study[43]		Probably total cost
400	50,000-100,000	Estimate by[47] I.S. Bowen		Probably total cost

* Figures given for large telescopes usually include such costs as land, site-development, building, and auxiliary instruments. Where the information was sufficient to eliminate these, the basic telescope cost is given.

† Scientific Research (November 1967), p 23.

Table 32

TRANSMITTING OPTICS – WEIGHT*

Diameter (in)	Weight (lb)
8	229-263
12	300
16	390-430
20	540
32	825
80	13,500
120	16,000

3.3 Transmitting Optics – Environmental Effects

Two effects of the space environment are particularly harmful to the transmitting optics and require careful attention. These are meteoroid erosion and damage and thermal deformations of the optics. They will be taken up in that order.

3.3.1 Meteoroids

Meteoroid effects may be divided for convenience into erosion by small particles and gross damage by larger ones. The problem has been given rather extensive attention, as the 311 references in a literature survey by Cosby and Lyle[50] attest. Even so, much of the work is not conclusive or does not permit direct design calculations. There has been a good deal of modeling, and measurements have been made on ballistic ranges at velocities somewhat larger than 10 km/s. However, it appears that this work cannot safely be extrapolated to the substantially higher meteoroid velocities (30 km/s and higher) which exist (at least near Earth).

Useful information on the probability of gross damage is obtained from direct measurements by the penetration and fracture sensors mounted on satellites.[50] Vanguard III, with four particle sensors, recorded no particles; Explorer VII, with one sensor, one particle; Explorer XIV, with four sensors, no particles; Explorer I, with one sensor, no particles; Explorer III, with one sensor, two particles; Midas II, with one sensor, no particles; and Samos II, with one sensor, recorded eight particles. Later, Explorer XVI, with eight kinds of sensors, in a 4-month period detected about 40 particles, allowing D'Aiutolo[51] to calculate impact rates. He obtained 10^{-5}particle/m^2 s for masses greater than 10^{-10} gram and between 10^{-5} and 10^{-6} particle/m^2 s for masses greater than 10^{-9} gram. Then, in 1965, Mariner IV recorded rates up to 3.3×10^{-4} particle/m^2 s for masses greater than 10^{-13} gram. More up-to-da·' information should be obtainable from Pegasus.

The salient feature of these results is that damage seems to be a relatively rare event, with some satellites sustaining no damage. Cosby and Lyle observed that the events detected are so sparse that meaningful average rates sometimes cannot be determined. Even the rates obtained for Explorer XVI seem to hold for very small particles near the low end of the mass range given. The sensors employed were intentionally made very fragile; they were designed to be penetrated by very small particles (evidently of the order of 10 microns in size). Mariner IV detected particles the size of dust (evidently down to about 1/4 micron in size). There does not seem to be clear evidence of even a single encounter with objects large enough to seriously damage equipment of ordinary size and strength, such as the transmitting optics. Thus it seems reasonable, pending further evidence, to discount the problem and suggest that no special protection (against gross damage) be provided. Like driving one's car, the hazards are real but not sufficient to deter. The performance of the optical systems on Mariners II and IV has led Becker to a similar conclusion[53]

Erosion by small particles also has received much attention, and again the direct applicability of much of the work is uncertain. However, two separate studies of the erosion of iron meteorities have yielded similar and quite encouraging numbers. Whipple and Fireman[54] found that the rate of erosion did not exceed 1.5×10^{-7} cm/year. Jaffe and Rittenhouse[55] obtained a rate less than 30Å/year, so that the results agree within a factor of two or less. Here the concern would be for the $\lambda/50$ reflecting surfaces, especially the primary, which would look straight into space. However, $\lambda/50$ is about 100 Å, so at least three years of erosion would be required to produce irregularities as large as the starting surface figure. Thus it would be years before any loss of gain would be noticed and decades before the reflector would be useless.

3.3.2 Thermal Deformations

It is the presence of air, as a heat-exchange gas, that largely reduces temperature differences on the surface of the Earth. The artificial production of low temperatures has, as its first requirement, the production of a hard vacuum. In space, this very condition results in large thermal gradients and hence distortions of the components and structures. This is one of the most severe problems in maintaining large, precise, optical systems in space, and in the past it often has not received adequate attention. There is a tendency to design sophisticated optical satellites which lack sufficient control of temperature and temperature gradients.

*These data were supplied by the Perkin-Elmer Corporation. The weights are for complete communications, telescopes, including parts such as secondary optics, beamsplitters, beampointing, and point ahead subsystems.

The reality of the problem with large optics is illustrated by experiences with the 200-inch Hale telescope, as related by I.S. Bowen of the Palomar observatory. "All of these mirrors we are talking about have such high thermal inertia and poor conductivity that it literally takes a week or two to get into equilibrium after a cold front passes. This is true whether we are talking of quartz or glass."[44] This is even more impressive when it is realized that, by the standards of the present study, the 200-inch unit is a relatively crude telescope. Its resolution is only about 1.5 microradians because it is far from diffraction-limited (for visible light). This is about the resolution of a diffraction-limited 1-foot aperture. The resolution of the 1-meter diffraction-limited aperture under consideration would then be about three times better (about 0.5 microradian).

Thermal effects on optics fall naturally into two categories: expansion of the structure and distortion of the elements. Expansion of the structure will be dealt with first. The nature of the problem and the appropriate solutions are illustrated by considering the critical matter of maintaining focus. (In an optical communications telescope, maintaining focus is equivalent to preserving beamwidth.)

To preserve the full resolving power of a 1-meter, diffraction-limited reflector of small f number (say, f/3) requires very accurate relative positioning of the elements. For visible light, the depth of focus of such an aperture is only several microns, while the focal length is three meters. Thus the distance from the primary mirror to a detector or to a secondary mirror must be maintained to about one part in 10^6. With ordinary construction materials, temperature could not vary by more than a fraction of a degree Centigrade.

One approach would be to use one material (such as quartz, beryllium, or aluminum) for the optical elements as well as for all supports. Then the entire structure would expand together; the change would be a matter of scale only, and focus would be preserved. Such a solution is not suitable for balloon flights (such as Stratoscope), where temperature changes rapidly, and probably not suitable for an orbiting telescope for the same reason. However, for a deep space flight this method might work out well. Temperature changes very slowly during the flight. However, near a planet there would be a rapidly changing thermal environment – akin to the Earth satellite situation.

Alternatively, the elements could be supported mechanically with very low expansion materials (or composite, thermally compensated structures showing very low overall expansion). Invar's coefficient is not so very low, and its ferromagnetism is usually objectionable in space vehicles. Pyroceram (Corning) and Cervit (Owens-Illinois) are more interesting. Over broad temperature ranges (tens of degrees), these can show a coefficient below $10^{-7}/°C$. It is not useful to give more exact data because these are families of materials and, to a considerable extent, characteristics can be tailored to need.[41] A very impressive candidate is Corning's new ultra-low-expansion (ULE) quartz, but it is available in rather limited forms and sizes and is of more interest for use in the primary mirror. It will be discussed in more detail later. There seems to be no good technical reason that large structural members could not be made of Pyroceram or Cervit.

The thermally compensated, composite structure represents an old idea. It has been designed in a new and suitable form by Rogerson, and it is referred to as the Rogerson tube.[56] It comprises re-entrant tubes of magnesium and Inconel, and its total change in length from -65° to -95°C is about one part in 10^6.

Still another means of maintaining focus is with active feedback, as is done in Stratoscope II.[57] It has been shown that this operates successfully during balloon flights. It may be required in general when large or rapid temperature changes are expected. Stratoscope II is similar in many respects to the design for an optical communications telescope proposed by Perkin-Elmer.[58] Fine adjustment of the optical axis of the telescope is accomplished by displacement of a transfer lens normal to the optical axis. The transfer lens is located where the eyepiece would be in an ordinary telescope. The error signal for positioning the lens is derived from an image divider in the last focal plane of the telescope. When focus is correct, a specified, small lateral displacement of the transfer lens produces the maximum rate of change of error signal, since the spot of light in the focal plane is then smallest. Thus, to focus the telescope, a small, intentional dither is applied to the lens while the secondary reflector is slowly moved axially. The secondary is then positioned where the largest rate of change of error signal is obtained. Such focusing is done only intermittently. In Stratoscope, this is under the remote command of the observer on earth, but automation of such a technique in space should be straightforward.

There are other positional tolerances, of course, but focus is about as demanding a condition as any. The other dimensions probably are readily maintained by passive means; i.e., by careful use of low-expansion materials. An example of one such necessary condition is collimation of the secondary mirror, but here the positional tolerance is much larger than for focus.

A tolerance problem peculiar to the communications situation (as opposed to ordinary telescopes which are receivers only) is maintaining the transmitter and receiver axes parallel. For a 1-microradian beamwidth, presumably the optical axes should be parallel to 0.1 microradian. This implies two telescopes spaced by supports wherein differences in expansion always are maintained smaller than the order of one part in 10^7. This is about an order of magnitude more demanding than focus, and even if very low expansion materials are used, it seems entirely unrealistic to attempt such an alignment in the thermal environment of space. (This does not imply that mechanical problems alone are not sufficient to preclude such an arrangement.) The evident solution is to employ common optics for receiver and transmitter, separating channels with dichroic beamsplitters. This is the approach taken by Perkin-Elmer.[58]

The second category of thermal effects on optics involves distortion of the elements. Thermally generated loss of figure of the primary mirror is probably the most difficult of all thermal problems. (The other telescope elements are much smaller and heat and cool more quickly, and there are smaller absolute differences of temperature across them.) The primary may be exposed directly to the sources of the disturbing radiation: the Sun and the Earth or planet. For other reasons as well, it is imperative that the telescope never point directly at the Sun, but it will often be impossible to keep sunlight from falling obliquely into the telescope tube. There will be a similar problem with reflected sunlight and thermal radiation from the planet, when the vehicle is near it. The tube interior would be as absorbent as possible, but reradiation will propagate energy to the mirror, and in general it will not be uniform across the diameter.

The third condition is that of a thermal gradient normal to the mirror axis. Then one side would expand more and have a larger focal length. Surface figure then would be lost, and this could not be corrected by focusing. For a 1-meter mirror of fused quartz, a 4°C difference between sides would be unacceptable.

Finally comes the most general case of a gradient both along and across the mirror axis. If the temperature difference between front and back differs from one side of the mirror to the other by more than 0.1°C, the surface figure will be spoiled. The difficulty is that fused quartz, which now is the preferred mirror material, has low thermal conductivity and thermal gradients of this magnitude are bound to occur unless very special precautions are taken.

Danielson[59] has experimented with techniques for alleviating this situation. When a heavy, blackened aluminum plate is placed just behind the mirror, it artificially increases lateral thermal conductivity and reduces temperature differences between sides by a factor of 10. Once this is done, it is possible to control the uniformity of the front-to-back gradient. It then is necessary only to keep the non-uniformity of the radiation coming down the telescope tube within reasonable bounds. Danielson has found that it should be sufficient merely to have a tube about two or three times longer than the diameter of the mirror.

Danielson has given a treatment of this, and he has outlined the situation for four possible thermal conditions of the mirror.[59] First, the entire mirror may be heated uniformly to the same final temperature. The effect of this is only to increase the focal length in proportion to the change in mirror dimensions. This could be handled readily by the type of active control of focus described before. The second condition is when there is a thermal gradient only along the mirror axis; i.e., the front and back surfaces each had its own uniform temperature. To a good approximation, the effect of this also is only to change focus; the front surface would curl but would remain a good paraboloid.

Recently, Spitzer and Boley published a thorough analysis of the thermal deformations in a satellite telescope mirror of fused quartz.[60] They considered the situation for a low orbit (below 800 km) satellite, since there the thermal environment is worst. (If thermal deformations can be handled in a low orbit, then diffraction-limited performance should certainly be attainable in higher orbit or in space, where the thermal environment is more uniform.) They concluded that, if the telescope tube is substantially longer than the mirror diameter and if the sunlit Earth does not shine directly on the mirror at any time, the likely distortions are certainly within optical tolerance for a 1-meter telescope and probably also for a 3-meter telescope. The situation would be similar for a system near Mars or Venus.

Thus it appears that careful engineering with known technology can lead to diffraction-limited performance of a 1-meter optical transmitter in space. However, the procedures required for doing this impose considerable constraints on instrument pointing. The primary mirror, at the bottom of a telescope tube "substantially longer" than the mirror diameter, must not receive radiation directly from the sun or a nearby planet (e.g., Mars). Clearly, the tube should be as long as is consistent with the size of the vehicle. A length of tube much greater than 10 feet probably is not realistic. (Generally, the optical design tends toward short, low f/number systems.) Hence, the primary diameter is, say, 1/3 the tube length, and the half-angle of the field from which the sun or planet must be excluded is 18.4 degrees. This would seem to exclude considerable regions of space from use by the optical transmitter. If the satellite were near the planet, there would then be a long effective occultation. This situation seems burdensome, and consideration should be given to erecting a long collimating tube outside the main vehicle after it is in space. The tube could be designed similarly to the present space unfurlable booms.

There is a real possibility of improvements through use of techniques and materials not yet so well established. Chief among these is the use of materials which might cause less distortion because they possess either a smaller thermal coefficient of expansion, a smaller specific heat, or a larger thermal conductivity. In recent years, these novel possibilities have received wide attention (References 44,56,57,61,62). Fused quartz, however, is an old, known, and proved material. Large diffraction-limited optics are costly in time and money, and it has been correctly pointed out that a cautious posture is essential.[44,63] Although no alternative to quartz is yet known well enough to justify its immediate use, this situation might soon change.

Metals have been considered. Aluminum is interesting because it is light and has large thermal conductivity.

However, it is very difficult to get a good polish on aluminum, and its long-term mechanical stability is uncertain.

Beryllium is exceedingly rigid; its ratio of modulus of elasticity to density is five times that of quartz. It has been used successfully for small mirrors, but there are numerous technical problems in its handling and stability that are not well understood. Thermal shock can spoil the surface of a good beryllium mirror, and its long-term stability is in doubt. Some measurements have shown considerable mechanical hysteresis. (By comparison, fused quartz has for years been known for its exceedingly low hysteresis. Schwarschild noted[64] that the 36-inch quartz primary of Stratoscope II returned to λ/50 after the temperature cycle of the balloon flight and the shock of landing.)

The vitreous ceramics are of interest. These are partly devitrified glasses, so they are 2-phase materials and their stability is not certain. For example, they may contain large internal stresses, and in very large blanks there is the possibility that the material may have several different local thermal coefficients. Nevertheless, it is possible to obtain material with a coefficient below 10^{-7}/°C in a specified temperature range. Meinel reported[44] that a 16-inch mirror was finished from Pyroceram (Corning's product) and cut in two without impairing its surface figure, thus demonstrating negligible internal stress. Both Meinel[44] and Dietz and Bennett[62] have reported that Cervit, Owens-Illinois' vitreous ceramic, has successfully been polished to surfaces that show sufficiently low scattering to be acceptable for optical mirrors. Thus the vitreous ceramics appear to be important possible substitutes for fused quartz in large optics if reduced thermal distortion is desired.

Quite recently, another contender has appeared and seems to show sufficient promise to merit first consideration among the novel materials. This is Corning's ULE (ultra low expansion) fused silica, number 7971. It appears to be doped (perhaps with TiO_2), and the resulting material evidently is similar to ordinary fused quartz except for its ultra-low thermal coefficient of expansion. Data supplied by E.T. Decker of Corning indicate an average coefficient of approximately $3x10^{-8}$/°C over the range-50° to +20°C. In addition, there is a broad region at about -10°C, and another at about +80°C, where the coefficient appears to be sensibly zero. Thus the material is at least an order of magnitude better than ordinary fused quartz, and perhaps far better than that.

The other properties of ULE fused silica at present seem to be good. Decker said that Corning's first (and worst) piece was sent to Meinel at the University of Arizona. It was 16 inches in diameter and 3 inches thick, and it showed large internal stress. Nevertheless, the optical shop at Arizona finished it to a surface as good as that obtainable on ordinary quartz. The long-term stability and behavior of ULE fused silica with temperature-cycling of the material are not yet known. ULE fused silica is grown from the vapor phase, and a boule 60 inches in diameter by 4 inches thick is obtained. It can then be sagged to an approximate parabola of small f-number. This avoids hogging out material and makes the best use of the 4-inch thickness.

In summary, Spitzer and Boley[60] have shown that the thermal problem for a λ/50 1-meter reflector in space can be solved, by very careful design, using even ordinary fused quartz. Novel materials, and ULE fused silica in particular, offer the hope that maintaining such a reflector may become quite easy. An important question is whether use of ULE quartz would relax the requirement that the Sun or (near) planet must never directly irradiate the primary mirror. This possibility deserves early consideration.

4. BEAM POINTING

Optical frequencies are considered for the carrier of deep space communications because it is possible, in principle, to generate narrower beams with smaller antennas that can be done with microwaves. The hope is that, at astronomical ranges, the narrower beams will yield an improvement in received signal-to-noise ratio and hence in communication rate. To what extent, in practice, this can be realized depends on a number of technical questions, but central to the problem is the matter of pointing very narrow beams. Since previously the need did not exist, the technology of pointing such narrow beams is new.

What pointing accuracies are required depends on how fine a beam is used. That in turn depends on the frequency and aperture selected, and such a determination lies outside this section. Therefore, a beamwidth will be considered that is limited by diffraction of the highest frequencies of interest from the largest aperture suitable for that frequency, since this is the most demanding situation: visible light (λ~0.5 micron) and a 1-meter aperture (following the argument given in Section 3). For these values, the width, from null to null, of the central maximum of the Airy diffraction pattern (far field of the uniformly illuminated aperture) is about 1 microradian. This beam must be pointed so it illuminates the receiver with about the central 1/10 of the beamwidth, so that the pointing accuracy is about 0.1 microradian. But more is required. Both pointing offset to compensate for velocity aberration (Bradley effect) and the position of the receiver must be determined to a similar tolerance. Finally, if the pointing is incorrect, there must be some means for generating a pointing error and making corrections.

The assessment of how accurately narrow beams can be pointed depends on whether the pointing reference is a coordinate system or a beacon from a source to which the beam is to be directed. The former case is more difficult and compounds the uncertainties in establishing the coordinate system, in pointing an optical axis with respect to it, and in fixing the receiver in the coordinate system. This is absolute, or open loop, pointing as opposed to relative pointing on a beacon reference.

4.1 Open Loop Pointing

The process of closing an optical communication link must begin with open loop pointing at one end. (Once the receiver at the other end has been illuminated, then the simplicity of pointing on a beacon reference can lead into an optical autotrack condition.)

The first requirement for pointing a fine beam is the generation of the beam itself. Until quite recently, diffraction-limited optics 1 meter in diameter were not available. Until then, there was no need for them: the atmosphere limited resolution, or beamwidth (see Section 3). Stratoscope II, however, rose far enough above the atmosphere (80,000 feet) to use the resolution of a 36 inch diameter diffraction-limited primary mirror. This called for techniques for finishing, and especially testing, the surface figure of large mirrors. Such a procedure, scatter plate interferometry, was suggested by Burch[65] and developed by R.M. Scott.[57] It was used to bring the Stratoscope primary to a paraboloid within a probable error of $\lambda/50$.[57,64] (This was the measured figure before and after flights, but it is not yet known how well it was preserved at altitude with a temperature of -50°C. This will soon be determined.[64])

With the means in hand for producing 1-meter diffraction-limited apertures, it is essential to determine how well a (visible) beam from such an aperture can be pointed on an open loop – or absolute – basis. If attention is restricted to the space vehicle or to a synchronous satellite (postponing for the time being the limitations set by Earth's atmosphere), it is remarkable how much technology attains angular accuracies of about 5 microradians (1 second) or somewhat better. This is true for the techniques of establishing the coordinate system, for stabilization of the platform upon which the pointing is done, and in setting an optical axis with respect to the coordinates.

Star trackers, for example, are now familiar devices for establishing coordinate systems on stars. Generally they have been designed for coarser accuracies, but there is no reason why they cannot specify a direction to within a fraction of the diffraction limit of their primary apertures. (Indeed, Stratoscope II, a 36 inch star tracker, can point to 0.1 microradian.) Small telescopes, 2 or 3 inches in diameter, can readily indicate a star direction to within 5 microradians.[56] An attractive arrangement would be an array of such units – perhaps as many as six – each pointing at a bright star and fixed with respect to each other by the angular separation of the stars. The variation in velocity aberration (an annual variation at the Earth) could be taken out by command to the pointing system, or, since it is changing slowly, the trackers could adapt to the shift in apparent star positions. The combined information would yield a coordinate system accurate to at least 5 microradians. [It is interesting that ITT Federal Laboratories (San Fernando, California) now offers a complete, small (3 by 15 inches) star tracker for which a tracking accuracy of 1-1/2 seconds is claimed. ITT Industrial Laboratories (Fort Wayne, Indiana) has photomultiplier tubes designed for star tracking. These are magnetically focused image dissectors and, in laboratory work, a tracking accuracy of about 0.7 second was reported.[66]]

Attitude stabilization of platforms which simulate space vehicles has been demonstrated at the NASA Ames Research Center, Moffet Field, California.[67,68] A two-rotor control-moment gyro was used for each of three orthogonal axes, and the platform was on a ball and socket air bearing. Attitude was maintained automatically to within 5 microradians.

If the coordinate system is better than 1 second and the platform is stable to 1 second, how well can an optical axis be pointed on the platform with respect to the coordinate system? Extensive optical experience suggests that this accuracy also is 1 second or better. Two companies make precision theodolites with which an optical axis can be set by scale readings with an uncertainty below one second. These are the DKM3 unit, made by Kern & Co., Ltd., Aarau, Switzerland, and the T-3 made by Wild Heerbrug, Ltd., Heerbrug, Switzerland.

Further technology relevant to this question was presented by D. Trumbo of Kitt Peak National Observatory.[44] A study was undertaken to improve the drive systems of telescopes 36 inches and larger. The polar axis could not be driven directly because of the large necessary torques, but the worm shaft could be. It appeared that by using digital techniques, the shaft position could be read out directly to correspond to an uncertainty in setting of the polar axis of about 0.05 second (1/4 microradian). There were larger pointing errors from the periodic error of the worm gear, but the control system itself appeared to be good to 0.05 second. After the Tucson Conference,[44] such a drive actually was tried on the Kitt Peak 84-inch telescope. In tracking accuracy and response to guide signals, it performed as expected. Since this method of telescope guiding uses digital readings of shaft position, it is as accurate when axis rotation is very small or zero as it is with large rates. It also is easily adapted to automatic tracking systems.

These comments need clarification when the optical axis being pointed is that of a really large Earth-based telescope. There will then be pointing errors induced by the dependence of gravity distortion of the structure on angle of elevation, by thermal gradients, and by errors of construction such as bearing runout and misalignment of the bearing axes. In the case of a 120 inch telescope, careful design can keep the accumulated error down to several seconds (perhaps 10 to 20 microradians).[69] However, it does not appear to be necessary to use such a large Earth-based telescope to initiate open-loop pointing. A large telescope would be wanted as the optical receiver and, if the receiver optics were used for the transmitted beacon as well (the common-optics of transceiver system), then these errors would be involved. However, a smaller separate telescope could serve as the transmitter, and for it such errors can be made smaller than 1 second (as in a precision theodolite). Furthermore, if the telescope is Earth-based, the atmosphere introduces even greater uncertainties, as will be seen below.

To summarize, it seems realistic to expect to point an optical axis with respect to a coordinate system (the stars in particular) with an overall accuracy of about 5 to 10 microradians when the vehicle is outside the Earth's atmosphere.

To settle the overall question of the accuracy with which a vehicle can be illuminated by an optical beacon needs specification of the uncertainty of determination of the vehicle position. This is done in Chapter 6, Section 1, where it is shown that after about 60 days of tracking information, which includes angular inputs, vehicle position can be known to better than about 2 x 10^4 feet, or 6 km. At a range of somewhat more than 10^8 km, which would be near encounter, the uncertainty in direction is only about 0.05 microradian. At smaller ranges it seems that as soon as accurate tracking information is available, on the tenth day after injection, the spacecraft is 2.7 x 10^6 km from Earth, and even there the uncertainty in position is only 3.7 microradians. It therefore should be possible to point an optical beam at the spacecraft with an overall uncertainty of about 5 to 10 microradians. This could be done from an Earth satellite. It would require a highly sophisticated satellite, but it probably could be done.

In many ways it would be simpler to point from Earth, except for one serious difficulty: random refraction by the atmosphere. This is the geometrical reciprocal of astronomical seeing, or loss of resolution. The transmitted beacon must be broadened, to the extent of the uncertainty in the angle of arrival of light from a star, to ensure illumination of the spacecraft. However, it is difficult to obtain data from astronomical observations which can be related directly to the determination of angular width of a beacon. Seeing usually involves integration of the angle of arrival over a short period of time – of the order of a second. Variations slower than that are taken out by telescope guidance. In general, variable-rate driving is used, and the telescope tracks the star. Since every effort is made to avoid loss of resolution, little is known about how great the absolute uncertainty can be over an entire night. Moreover, the good seeing which has been reported was obtained at night. The communications system must operate during the day as well, when random refraction would be worse. During the day the ground temperature is higher than the air temperature (just the reverse of the situation at night), and warm air will rise around the observatory, or ground station, causing turbulence and poorer seeing. Seeing also may be worse with decreasing elevation angle, but again the available information is inadequate. (The average refraction by the atmosphere for slant paths is not under consideration here. Not only is this rather well known and correctable to about 1 arc second for changes in average barometric pressure and temperature, but appropriate star sightings can provide automatic correction.) The overall situation has been described succinctly by Bowen: "Unfortunately our ignorance of quantitative seeing is very profound."

A good review of seeing is given by Meinel.[61] Data are reported for a variety of observations in good astronomical locations. A representative value for seeing disc diameter seems to be about 7 microradians (1-1/2 seconds). All the data, however, appear to involve an integration time of the order of a second or less. Meinel has indicated[70] that, at observatories, seeing of about 7 to 10 microradians usually is maintained for extended periods. However, there are occasional periods of very bad seeing, such as when a cold front passes. Seeing then can "blow up" to 100 microradians. S. Vasilevskis has commented that, when the temperature has dropped at Lick Observatory, the seeing is so bad that there is no use in attempting to avoid thermal distortions of the mirror.[44] The total meandering of a star image on an ordinary night has been observed at Bell Telephone Laboratories (an ordinary location in New Jersey), and an overall uncertainty in angle of arrival of about 140 microradians was found.[71] Various observers have found that laser beams, propagated over short (a few hundred feet or more) horizontal paths near the ground, diverge because of atmospheric refraction typically by 50 to 100 microradians.

It is true that the location for a ground station would be selected with great care to avoid the worst conditions of random atmospheric refraction. But it is also true that selection would be constrained by probability of cloud cover, that a large number of stations around the world would be needed for sufficient diversity, and that selecting even a conventional observatory site is now not a simple matter for astronomers.[61] Mountain top locations, which are best at night, might be bad in the day. Air heated by the mountainside in the daytime results in strong vertical convection at the top. To make matters even worse, this tends to result in daytime formation of clouds above the mountains. If the daytime seeing problem is caused largely by rising air which has been heated by the ground, then good daytime locations might be over water. Water in a lake has a large effective thermal mass, owing both to the large depth of absorption and to convection. Thus the diurnal variation in the surface temperature of water is much smaller than that of land, and the effect on the air over the water should be correspondingly reduced. Possible locations would be on small, flat islands or peninsulas, or, perhaps better still, on man-made platforms elevated well above the water on stilts.

The astronomers who study the fine structure of the Sun, such as sunspots, have a similar site requirement for best daytime resolution. According to Becker[72] there is a belief among sunspot astronomers that a large body of water adjoining the site is helpful. Becker reports[72] that the quality of daytime seeing is dependent on location. At Sacramento Peak, 3 arc seconds is representative, with 10 arc seconds being the worst condition. However, at a site used by Becker in Australia, the worst condition was 20 arc seconds.

It is concluded that it is unrealistic to use values ordinarily cited for astronomical resolution for the assessment of the adequate width of an Earth beacon. Not enough is known about how severe random refraction might be, and the spacecraft must be illuminated by a certain power density with very high probability, day or night. Known worst conditions indicate that the beacon width probably should be 50 to 100 microradians or perhaps more. However, it is possible that the right station in the right location (perhaps elevated over water) would give better overall seeing. This is an area that very much needs attention, and a funded study is recommended.

4.2 Pointing on a Beacon Reference

Once one end of the communication link has been illuminated by a beacon at the other end, then the technology of accurate pointing changes in kind and degree. More accurate pointing is then accomplished in a far simpler way. If a common-optics transmitter-receiver is used, then pointing the transmitted beam is only a matter of tracking the beacon by the receiver. That is, there is no reason why pointing accuracy cannot be made as good as the accuracy with which an optical system can track a star. As discussed above, this accuracy is a fraction of the diffraction limit of the aperture of the telescope. This is true provided only that the received signal is strong enough to give good statistics in the tracking system.

Demonstration that this can be done was provided in the laboratory at the Perkin-Elmer Corporation.[49,58] A tracking accuracy of about 0.5 microradian was obtained with a 40-cm primary mirror. This amounts to about 1/8 of the full (null-to-null) beamwidth of such an aperture. (Full details of this work are available in Perkin-Elmer reports to NASA.[58]) So far, this has been done with very strong beacons (the full beam from a laboratory laser), but the plan is to develop the system at progressively lower signal levels.

Additional directly relevant experience was obtained with Stratoscope II. Its tracking system can, in the main, be used in optical communications telescopes. Stratoscope II was lifted by balloon to about 80,000 ft and there the entire 2-1/2 ton telescope body pointed to a star with an accuracy of about 5 microradians. This amounted to coarse pointing – the equivalent of stabilization of a platform in space. (It is interesting that this agrees so well with the laboratory work on platform stabilization at NASA/Ames.)[67,68] In the successful flights, the transfer-lens system for fine pointing was not incorporated. However, this system has been simulated in the laboratory,[73] and it seems [illegible] that this subsystem will enable Stratoscope II to track stars to within 0.1 microradian[64] (which is about 1/10 the Stratoscope II beamwidth).

This success with Stratoscope II rests on some reasonably well-defined technical conditions. It will be instructive to examine these. Tracking on a star or optical beacon requires that the received photon flux be large enough to give good statistics within the period of the highest frequency disturbance which has sufficient amplitude to misalign the optical axis of the tracking system. Since pointing must be maintained to within, say, 1/10 of the beamwidth, then disturbances causing misalignments of 0.1 microradian would be objectionable in a 1-meter (microradian) system. The most difficult point to assess is the level and spectrum of mechanical disturbances to be expected in the spacecraft, but Stratoscope II offers some guidance.

Stratoscope II employed a mercury-float for an azimuth bearing and flexure bearings for the fine elevation and roll bearings. The flexure bearings had a maximum excursion of ±5 degrees: beyond that, ball bearings provided larger rotations. Restoring torques between supporting frame and telescope body were provided by magnetic torque motors. Thus, while the telescope was actually tracking, there was no mechanical motion whatever except for rotation of the mercury float and bending of the flexure bearings. The forces guiding the telescope were provided magnetically. Schwarzschild[44] pointed out that, for tracking faint stars, it is essential that there be no moving parts of the ordinary sort in the vehicle. This specifically excludes inertial wheels, gas jets and valves, ball bearings, and gears. Thus, for example, the only acceptable bearings are flexure bearings.

Information regarding the minimum photon flux can be obtained from the weakest star which could be tracked with such a system. This was a ninth magnitude star, and tracking was limited by the presence of significant mechanical disturbances up to a frequency of about 10 cycles. In the Stratoscope situation, the background light and detector dark current were not dominating noise sources; the important noise was quantum noise of the star, or signal, photons. The detectors were RCA 7265 photomultipliers, having an S20 photocathode. This photocathode has an average quantum efficiency of about 0.1 over about a 2100 A range. The overall optical loss was about 10 dB. Therefore, the rate of detected photoelectrons was about 10^5 per second, or 10^4 per cycle of 10 cycle disturbance. The important point to retain is that, were there conventional moving parts in the telescope, then the photoelectron rate would have had to be much higher than 10^5 per second.

It is difficult to project this system directly to the spacecraft situation, and especially to estimate the amplitude and spectrum of disturbances. However, some general observations can be made. The communications telescope should be designed so that while it is tracking (or communicating) there are no parts in motion other than small elements, such as a transfer lens, mounted on flexure bearings. Since the main vehicle must have conventional attitude-contr l devices, such as control moment gyros and gas jets, the telescope must be decoupled from the main vehicle through a soft gimballing or the like. It can be argued that the environment of deep space is mechanically more benign than that of a balloon slowly swinging in the upper atmosphere. However, the main vehicle of the spacecraft would provide an environment which may well be mechanically much noisier than that of a passive balloon. Despite mechanical decoupling, it will be very difficult to keep low-frequency disturbances out of the communications telescope. Laboratory work with sensitive instruments shows that isolation of large disturbances below about 10 cycles is quite difficult. It may be, then, that the beacon strength would need to be large enough to allow tracking in the presence of a significant amplitude of disturbances up to 10 cycles in frequency.

It is apparent that knowledge in this area is very inadequate and that careful measurements are needed. It is recommended that the type of communications telescope that was developed by Perkin-Elmer be placed on a platform which can simulate the space vehicle, such as the facilities at NASA/Ames, and acquisition and tracking studies be performed. The laser beacon should be reduced in intensity to the signal level expected in deep space. By means of a beamsplitter, background optical noise should be superimposed on the beacon to simulate the presence of Earthshine. Finally, the noise sources expected on the spacecraft (control moment gyros, gas jets, and any motors, gears, and bearings) should be operated, and the decoupling which the telescope then requires should be studied.

4.3 Pointing Offset

Due to large relative velocities between Earth and spacecraft, it is necessary to offset the axis of the transmitted beam with respect to the axis of the tracked beacon because of velocity aberration. This is a serious matter because the offset can be orders of magnitude larger than the beamwidth itself. The relative velocity normal to the line of sight can be as large as approximately the Earth's orbital velocity around the Sun, which is $10^{-4}c$, where c is the velocity of light. Since

$$\theta_{\text{Offset}} = 2\frac{V_n}{c}$$

where V_n is relative velocity normal to the line of sight, then θ_{Offset} can be approximately as large as 2×10^{-4} rad, or 200 microradians. This is 200 times a beamwidth of one microradian (still taking the "worst case" of visible light and a 1-meter aperture), or 2000 times the allowed error in pointing.

Whether a correct pointing offset can be executed depends on whether tracking information yields sufficiently accurate values for V_n and on whether an optical device can be developed to generate the offset with sufficient accuracy. Chapter 6, Section 1 shows that after about 75 days of tracking, the velocity of the spacecraft can be known within about 0.002 ft/sec. Therefore the error in the value for pointing offset is only about $2 \times 2 \times 10^{-3}/10^9 = 4 \times 10^{-12}$ radian, or 4×10^{-6} microradian, where the velocity of light is 10^9 ft/sec. Since the required accuracy is only 0.1 microradian, the information is far more than good enough.

The means for producing the offset is somewhat more difficult. In the Perkin-Elmer communications telescope,[58] thin circular wedges (a Risley prism) were rotated to vary both magnitude and direction of offset of the transmitted beam only. This type of device has the advantage of large demultiplication of error in the rotational setting of the disks from the error in the offset of the refracted beam. It was demonstrated that the prisms could produce, by remote control, an offset which was accurate to 0.5 microradian. As it is, this is about five times poorer than is required for offsetting a microradian beam. Moreover, this particular device could produce a maximum offset of only 30 microradians. In the spacecraft an offset perhaps seven times larger would be needed. Thus wedges about seven times thicker would be needed, and the offset would be seven times more sensitive to errors in the rotational setting of the disks. There also is the fact that the offset at the wedge is larger than the offset outside the telescope by a factor which is the telescope magnification. In the Perkin-Elmer system this gave a maximum internal offset of 10 arc minutes, so that the equivalent unit on a spacecraft would need a maximum internal offset of more than a degree. It is not now clear whether this would raise subsidiary optical problems such as aperturing or aberrations.

Finally, there is the difficulty that this offset system requires noisy moving parts. The disks are driven by gear trains and servo motors. As discussed above (Section 4.2), they should be excluded from a system designed to track on weak beacons. The presence of such mechanical disturbances would increase the necessary beacon power greatly, and already the beacon power requirement is severe

(Chapter 4, Section 4). A possible alternative, which should be given close attention, is to offset by a pair of crossed electro-optical prisms. The refraction produced in two orthogonal directions then would be varied simply by varying the strength of the electric field applied between the parallel faces of the prisms. Using $LiTaO_3$ in the prism and a prism angle of 35 degrees, a deflection of 100 microradians could be produced by an electric field of 3000 volts/cm. Previous work has been performed at Bell Telephone Laboratories and elsewhere on single prism devices of this type, but so far little has been done with two-prism, two-dimensional deflectors. There are difficulties not yet resolved with such devices, for example, changes in optical beam polarization due to deflector birefringence, but the prism system remains a possible solution to the very objectionable noise in mechanically operated beam deflectors. Appropriate device development is recommended.

4.4 Pointing Error

As was seen above, there are important uncertainties in the pointing of the spacecraft transmitted beam. It is by no means certain that all of the necessary technology can be developed for pointing with the necessary accuracy – especially for generating the correct offset. Moreover, there is the possibility of the steady development of systematic errors, equipment malfunction, etc. It is necessary to raise the question of what must be done if the Earth station fails to receive the spacecraft beam, even though it should, or if the Earth station receives the signal, but it becomes steadily weaker. That is, it seems appropriate to plan to incorporate a method for generating a pointing error for the spacecraft beam, provided this can be accomplished without an excessive penalty in on-board complexity and weight.

Several methods can be used for measuring the spacecraft beam pointing error. One is an array of secondary receivers surrounding each Earth station. In principle, the amount of error could be determined by comparing the signals received at the elements of the array. There are several disadvantages to such a system. First, if the Earth stations are on the ground, this arrangement would greatly increase the already large number of stations needed for diversity. Even at that, the system could operate only in the acquired condition, when the signal is received but the beam is not centered on the station. When the error is large and the beam is not acquired, no information on pointing error can be obtained. Finally, if the stations are on the ground, local atmospheric conditions will cause signal fading peculiar to the various secondary receivers, and spurious pointing errors will result. If it is placed in synchronous orbit, the array can be only one-dimensional, and errors out of the equatorial plane cannot be read.

Another possible method would be to structure the spacecraft beam. The wings of the beam, for example, could be coded according to quadrant. In some respects this is an improvement over the receiver array method, but some optical complexity must be added to the spacecraft. The principal shortcoming, however, is that again there is no error information when the beam is not acquired by the Earth station.

A method which would function in both the acquired and not-acquired modes is scanning of the spacecraft beam. In the acquired condition, the beam could be scanned, for example, in a square pattern, with each corner or each side of the square coded. Scan amplitude would be only a fraction of the beam radius so that when there is a pointing error the modulation of the received signal would be no more than, say, 20 percent of the on-axis intensity. When pointing is correct, there would be small modulation (at the fourth harmonic), and the signal would be slightly less than the on-axis maximum. When the signal has not yet been acquired or has been lost, this same system could be used, where now the beam diameter and scan amplitude would merely be increased until, using long integration times, the signal is reacquired. (This scheme is elaborated on in Chapter 4, Section 3.3, and in Appendix 5.)

Another important advantage of the scan technique is that little complexity needs to be added either to the ground station or to the spacecraft. Indeed, the spacecraft communications telescope already incorporates the optics needed to generate such a scan: the prism system (Risley, electro-optic, or otherwise) required for pointing offset. The scan would be executed about the pointing offset as a mean position. (If the scanning operates steadily, as here proposed, it militates all the more for a nonmechanical pointing-offset device.) It would seem that only the spacecraft circuitry would need to be augmented for addition of the scan capability.

The frequency of scan must be chosen with care because of atmospheric effects. Random atmospheric refraction causes an amplitude modulation of star brightness. This is most severe with small apertures (producing twinkling to the eye), but even for large apertures it would impose noise for all frequencies within the characteristic spectrum of the fluctuations. Many observations have been made on the spectra of star scintillations (for example, see Meinel's article on seeing)[61] and on the fluctuations in the received signal when laser beams are propagated through the atmosphere. In all cases the widths of the spectra are of the order of hundreds of cycles. The greatest range of spectral width, at 0.63 micron wavelength, was reported by Subramanian and Collinson,[74] who found that spectral width is dominated by atmospheric conditions. Those measurements are summarized in Figure 82. The broadest spectrum, measured during a rain squall, was about 1000 Hz wide. It appears, therefore, that a scan frequency somewhat higher than 1000 Hz would be a good choice

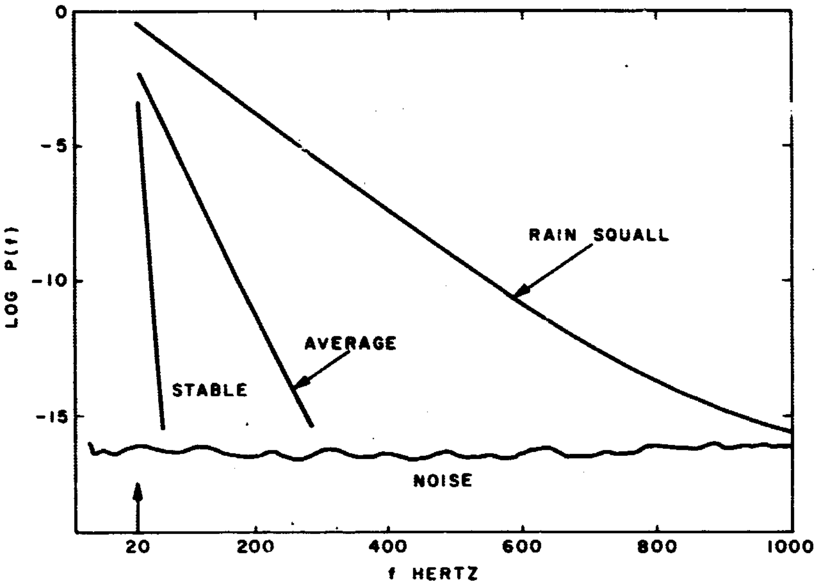


Figure 82. Modulation power spectrum, P(f) vs. f, induced by atmospheric turbulence on a 0.63 micron laser beam (P(f) = constant $\cdot e^{-\alpha f}$ where α is a constant dependent on weather conditions)

in the acquired condition. Atmosphe.ic modulation should not then be significant. In the not-acquired condition, larger beams and lower data rates probably would require much slower scan rates.

If it is possible to develop a suitable pointing-offset subsystem, then it appears feasible and appropriate to design the supplementary pointing-error correction system. A large gain in communication reliability can be obtained in this way at relatively small cost in complexity.

5. RECEIVING OPTICS

For the purpose of this study, receiving optics differ according to whether they ·e suitable for coherent reception or simply are collectors of optical power and can be used for incoherent reception only.

5.1 Coherent Receiving Optics

For a receiver to qualify as coherent, it should be, from an optical viewpoint, diffraction-limited. This implies simply that an optical wave which is spatially coherent will pass through the rec··er without significant distortion of the wavefront. Such a wave will show a divergence in the far field limited only by diffraction, and the optical system which transmitted it is then diffraction-limited. The detection system of the receiver can then make use of the interference effects which plane waves will allow. The effective area of coherent receivers can be increased by using arrays of collectors, each smaller than the coherence area, and combining the outputs coherently (see Chapter 5, Section 5.3 for a brief discussion).

To be diffraction-limited is a stringent requirement for large-diameter optics. Such surfaces are difficult to generate and very difficult to maintain. Thermal effects and gravity are extremely serious perturbations, and most investigations of optical communications have not tended to give due weight to these difficulties.

As in Section 3, the condition of the primary mirror is the main consideration. The primary must have an rms deviation from the correct shape (usually a paraboloid) not more than 1/50 wavelength of the radiation used. As discussed earlier, large telescopes are not this good for visible light, but most of them are good for a wavelength of 10.6 microns (or are close to such a figure). This is the case, for example, for the 84 inch mirror on Kitt Peak[44] and the 200 inch Hale at Palomar.[75] Also, as discussed above, technology has now been developed to produce mirrors which have 1/50 wave surfaces in the visible for sizes up to a 36 inch diameter (Stratoscope II).

The cost of telescopes is discussed in Section 3.1, and the data are summarized in Figure 81. The cost of a communication receiver should not differ much from these values since they already include dome and mount. Perkin-Elmer has performed a cost study[76] for a 120 inch receiver (diffraction limited at 10.6 microns). Their $4 million figure fits the other data in Figure 81 well.

The most serious and most ignored problem for a large Earth-based optical receiver is that of thermal perturbations. Bowen's statement[44] that large telescope mirrors literally take a week or two to get back into equilibrium after a cold front passes bears repetition. Since the problem then is a difference in cooling rate from the different sides of the mirror, the practice (with the 200 inch) is to start 12 fans circulating air rapidly through the support structure in the hope that the structure will change about halfway to the new ambient temperature in an hour or two. One Hartmann test showed that a 5 degree drop in temperature caused the outer 25 to 35 cm of the mirror to curl back by 0.3 arc seconds, causing an image spread of 1.5 arc seconds.[44] This is very serious for astronomical purposes, since the mirror figure is then five times worse than it should be. It would also be very serious for coherent optical communications at 10 microns, since such a thermally deformed surface is far from diffraction-limited at 10 microns.

Various measures are taken by the astronomers. In addition to using fans, insulation such as bags filled with aluminum foil is added to the edges and back of the mirror. It is clear that the mirror must be quartz, but even this is by no means sufficient. The astronomers make it quite plain that thermal distortion is an unresolved problem[44] even under night-time observatory conditions.

The environment for ground-based optical communication receivers is far more severe and far more uncertain, since operation is required during the day as well. Observatories obtain their good seeing and (normally) uniform thermal environment by working at night on high ground. The ground is colder than the air at night, and the chilled, mixing air drains off downslope. By erecting a suitable building and elevating the telescope well above this turbulent, mixing region, a quite uniform environment is obtained for the telescope. In the day, however, the warmer ground causes an updraft and mixing air steadily rises around the telescope dome. The telescope then is also exposed to a heavy thermal load produced by thermal radiation from the sky and scattered sunlight. The most sanguine extrapolation from nighttime experience suggests that the thermal distortions of the mirror will be very great during the day. How great is not now clear, but it is equally unclear whether coherent (i.e., diffraction-limited) reception at 10.6 microns will be technically feasible at all. To make matters even worse, the diurnal variation in this thermal load must be contended with. The most serious condition is when outside temperature changes rapidly, particularly when it drops. Yet, for much of a typical mission, it would be necessary to open the dome and start "observing" just before sunset and on into darkness – a thermal environment which changes profoundly in hours.

Note in particular the mission displayed in Figure 134, Chapter 6. At encounter with Mars (when the wide-band link would be most needed), the Earth station must operate at just this time – before sunset and into nighttime.

Telescope protection of the type now used for observatories almost goes without saying. A dome which opens to a limited part of the sky is needed not only in an attempt to control the thermal environment but also to protect the telescope against dirt and wind. The dome and telescope should have independent foundations. Air-conditioning of large capacity and versatility would be needed in a further effort to match the dome conditions to outside conditions, although it remains uncertain that thermal degradation could be eliminated. Another serious thermal load is the dissipation from the communications equipment, and especially from the laser beacon if the receiver is to be a common-optics transmitter as well. With the efficiencies and output powers sometimes contemplated, tens or even hundreds of kilowatts might have to be removed.

Another problem, less serious but still requiring attention, is variation in gravity distortions when the optics are pointed in different directions. This influences the nature of the mount for the whole telescope as well as the nature of the primary mirror mount in the telescope. Most astronomical telescopes have equatorial mounts in which one rotation axis is made parallel to the Earth's axis of rotation, and the other axis is perpendicular. Rotation about the polar axis alone then keeps the optical axis pointed approximately at the stars. However, in all its varieties, the equatorial mount involves a relatively large mass (including the hardware which produces the polar rotation) whose gravity loading changes with rotation. A preferred choice would be the type more common in radio telescopes: an elevation bearing held in a vertical yoke which itself turns in an azimuth bearing. With this arrangement, rotation of the aximuth bearing generates no change in gravity deflections. This was described in some detail in the study on a 120 inch receiver.[69]

Proper mounting of the primary mirror is a massive mechanical problem. The 200 inch Hale telescope has an intricate system of 36 counterweights. Even so, the extensive discussion of the situation at the Tucson conference makes it apparent that gravity distortions are still a serious problem.[44] For example, Bowen pointed out that testing of the mirror is meaningless unless it is performed in the telescope. To be sure it is properly adjusted, a large mirror must be tested in all positions. Some new technology which should be followed closely is a pneumatic support system for the mirror. This has a mercury-filled neoprene tube inside the mirror cell, around the entire edge of the mirror. When it is on edge, the mirror floats on mercury, and, when it is reclining, it is supported by hydrostatic air pressure. Hoag[44] has shown the improvement in mirror figure that can be obtained by using such a support system.

5.2 Incoherent Receiving Optics

5.2.1 Reflector Tolerances for Incoherent Optical Receivers

Incoherent optical receivers (see Chapter 4, Section 7) may employ large nondiffraction-limited collecting apertures. Indeed, some studies[77] have suggested that appropriately coated millimeter antennas might be used for this purpose. It is pertinent to determine the tolerances required on large optical collectors for use in incoherent receiving systems. See also Appendix 3, which contains a general discussion of antenna tolerances. In this section, known results[78] on the behavior of lenses or mirrors with random imperfections are applied to determine the relationship between mirror size and "roughness" statistics, detector size, and field of view (to background radiation) for incoherent optical receivers.

5.2.1.1 Relation Between Geometric Imperfections and Phase Error. Assume for definiteness that an imperfect mirror is used as a collecting element; the corresponding discussion for an imperfect lens would be quite similar. Let the departure of the mirror surface from the ideal parabolic shape be denoted by m(x′, y′) where x′, y′ are transverse rectangular co-ordinates in the plane of the mirror. Then assume the mirror distortion is a two-dimensional stationary Gaussian random process with isotropic statistics. The distortion has covariance.[78]

$$\begin{aligned}&\underset{\sim}{\varphi}_m(\tau_x, \tau_y) \equiv < m(x' + \tau_x, y' + \tau_y)\, m(x', y') > = \\ &\underset{\sim}{\varphi}_m(\tau), \tau \equiv \sqrt{\tau_x{}^2 + \tau_y{}^2}\end{aligned} \tag{14}$$

The corresponding spectral density is

$$\begin{aligned}&\underset{\sim}{P}_m(f_x, f_y) = \underset{\sim}{P}_m(f) = 2\pi \int_0^\infty \underset{\sim}{\varphi}_m(\tau) J_0(2\pi f \tau)\tau d\tau, \\ &f \equiv \sqrt{f_x{}^2 + f_y{}^2}\end{aligned} \tag{15}$$

with a similar inverse relationship. A complete description of the statistics of the geometric imperfections of the mirror is now given by specifying either the spectral density $P_m{}^{(f)}$ or equivalently the covariance $\phi_m(\tau)$.

The results in Reference 80 are stated in terms of the statistics not of the geometric imperfections m of the reflector, but rather of the phase deviations γ of the reflected wave (just in front of the mirror) from the phase-front for an ideal mirror. Consequently, the phase statistics must be related to the geometric statistics, i.e., $\underset{\sim}{P}_\gamma(f)$ of Reference 78 must be determined in terms of $\underset{\sim}{P}_m(f)$ above.

If m(x′, y′) varies slowly compared to the free-space wavelength λ, this relationship is easily established by using a geometric optics approximation. A typical ray striking the mirror at x', y' travels an extra distance 2m(x', y'), suffering an additional phase shift

$$\gamma(x',y') \approx \frac{4\pi}{\lambda}\, m(x',y') \qquad (16)$$

Therefore

$$\underset{\sim}{P}_{\gamma}(f) \approx \left(\frac{4\pi}{\lambda}\right)^2 \underset{\sim}{P}_m(f) \qquad (17)$$

subject to the restriction that

$$\underset{\sim}{P}_m(f) \approx 0\,,\ f > \frac{\text{const.}}{\lambda} \qquad (18)$$

where const. is smaller than 1.

Furthermore, in the analysis of Reference 77 any components of γ having spatial fiequencies $f > 1/\lambda$ are not propagated to the focal plane, but result in only local fields in the immediate vicinity ($\sim\lambda$) of the mirror, similar to the evanescent fields present in cut-off waveguides or in total internal reflection. Therefore, if $\underset{\sim}{P}\gamma(f)$ had significant components for $f \underset{\sim}{>} 1/\lambda$, they should be dropped before using the formulas of Reference 76, i.e.,

$$\underset{\sim}{P}_{\gamma(\lambda)}(f) = \begin{array}{ll} \underset{\sim}{P}_{\gamma}(f)\,, & f < \frac{1}{\lambda} \\ 0 & ,\ f > \frac{1}{\lambda} \end{array} \qquad (19)$$

and the corresponding modified covariance[78]

$$\underset{\sim}{\varphi}_{\gamma(\lambda)}(\tau) = \underset{\sim}{\varphi}_{\gamma}(\tau) \circledast A_{\lambda}(\tau) \qquad (20)$$

where[2]

$$A_{\lambda}(\tau) \equiv \frac{\pi}{\lambda^2}\ \frac{J_1\left(\frac{2\pi}{\lambda}\tau\right)}{\frac{\pi}{\lambda}\tau} \qquad (21)$$

and $\circledast$ represents the two-dimensional convolution operator, should be used in the statistical analysis of Reference 78.

By using Equations (17) and (18), it is easy to compute $\underset{\sim}{P}\gamma(f)$ from $\underset{\sim}{P}_m(f)$ for f smaller than $1/\lambda$, and the value of $\underset{\sim}{P}\gamma(f)$ for $f \underset{\sim}{>} 1/\lambda$ is irrelevant. In the region of $f \sim 1/\lambda$ an accurate calculation appears difficult. (Furthermore, the small-angle approximations of Referer.ce 78 will limit the accuracy of the results far from the axis of the optical system, corresponding to this region.) Consequently, it is assumed:

$$\underset{\sim}{P}_{\gamma}(f) = \begin{array}{ll} \left(\frac{4\pi}{\lambda}\right)^2 \underset{\sim}{P}_m(f)\,, & f < \frac{1}{\lambda} \\ 0 & ,\ f > \frac{1}{\lambda} \end{array} \qquad (22)$$

in the following work. From Equation (20) the covariance of the phase deviations will consequently be

$$\underset{\sim}{\varphi}_{\gamma}(\tau) = \left(\frac{4\pi}{\lambda}\right)^2 \underset{\sim}{\varphi}_m(\tau) \circledast A_{\lambda}(\tau) \qquad (23)$$

where $A_{\lambda}(\tau)$ is given by Equation (21) and is discussed in Appendix B of Reference 78.

A precise investigation of the accuracy of this assumption appears difficult. It is known from the work of S.O. Rice[79] that this assumption is not strictly true; in particular, he points out that if the surface distortion is the sum of two sine waves of wavelengths less than λ, there will be scattering even though neither of the two would yield any scattering if it were present alone. For the case considered by Rice; i.e., rms deviation $\ll \lambda$ and rms slope $\ll 1$, it would be possible to investigate the error introduced by the assumption of Equation (22) in detail. Furthermore, this is an interesting case corresponding to a good mirror or antenna. However, other cases not covered by Rice's analysis will also be of interest here. Such an investigation will not be undertaken here; it will be assumed on intuitive grounds that Equations (22) and (23) give useful results. This question does not appear to have been considered in any of the standard treatments of random imperfections in antennas, lenses, or mirrors.

5.2.1.2 Statistical Model of the Reflector. A great variety of mirror distortion spectra $\underset{\sim}{P}_m(f)$ is possible, depending on the size, construction, and method of manufacture. The choice of $\underset{\sim}{P}_m(f)$ here will be directed toward the possibility of the use of a good millimeter wave antenna at optical frequencies. It is further assumed that the reflector surface is of good optical quality, i.e., mirror-like as for a searchlight reflector rather than a matte finish, e.g., dull aluminum. Although the reflector is assumed to be good at millimeter-wave frequencies, it will be assumed that it is poor at optical frequencies; i.e., the reflector is not diffraction-limited at optical frequencies, as a good telescope mirror would be. Even within these restrictions a wide choice of roughness spectra $\underset{\sim}{P}_m(f)$ is admissible, and two examples will be chosen as representative although others are clearly possible.

Gaussian Covariance. The first example assumes a Gaussian covariance and spectral density:

$$\underset{\sim}{\varphi}_m(\tau) = M^2 e^{-(c\tau)^2} \qquad (24)$$

where M = rms deviation of the mirror surface from the ideal paraboloid.

$$\frac{1}{c} = \text{correlation length of mirror distortion.} \qquad (25)$$

The corresponding spectral density is

$$\underset{\sim}{P}_m(f) = M^2\, \frac{\pi}{c^2}\, e^{-\left(\frac{\pi f}{c}\right)^2} \qquad (26)$$

A related quantity of interest is the rms gradient of the random surface $m(x',y')$, denoted by M'; for the Gaussian case [Equations (24) to (26)]

$$M' \equiv \sqrt{<|\nabla m|^2>} = 2cM \qquad (27)$$

Two assumptions are made in this case: (1) The mirror is a good antenna at $\lambda = 1$ mm $= 10^{-3}$ meter but a poor antenna at $\lambda = 1$ micron $= 10^{-6}$ meter. (2) The spectral density of the mirror distortion has fallen off to essentially zero at $f = 1/\lambda$ for $\lambda = 1$ mm $= 10^{-3}$ meter.

The first assumption requires

$$10^{-6} \ll M \ll 10^{-3} \text{ meter} \qquad (28)$$

so that

$$\tfrac{1}{2} \times 10^{-5} < M < \tfrac{1}{2} \times 10^{-4} \text{ meter} \qquad (29)$$

A reasonable millimeter-wave antenna will probably have M closer to the upper limit of Equation (29) than to the lower, so that

$$M \sim \tfrac{1}{2} \times 10^{-4} \text{ meter} \qquad (30)$$

in the following.

The second assumption yields

$$\lambda c < 1, \ \frac{1}{c} > \lambda \text{ for } \lambda = 10^{-3} \text{ meter} \qquad (31)$$

consequently,

$$\frac{1}{c} > 10^{-3} \text{ meter} \qquad (32)$$

stating that the correlation length is greater than 1 mm. This condition [Equation (32)] insures that truncating $\underset{\sim}{P}_m(f)$ of Equation (25) at $f = 1/\lambda$ for $\lambda = 1$ mm $= 10^{-3}$ meter will reduce the mean square mirror distortion M^2 by less than 0.0054 percent, and reduce its mean square gradient M'^2 by less than 0.058 percent; these numbers correspond to the minimum correlation length $1/c = 1$ mm $= 10^{-3}$ meter, and will decrease rapidly as the correlation length increases. The corresponding reductions would of course be far smaller for $\lambda = 1$ micron $= 10^{-6}$ meter. Consequently the resulting modification in the form of $\underset{\sim}{\phi}_m(\tau)$, represented by the operation of Equation (23), may be neglected for small τ, which proves useful in subsequent applications of the results of Reference 78 both for millimeter-wave and for optical frequencies. Equations (31) and (32) and (29) and (30) when substituted in Equation (27) further guarantee that the rms gradient of the mirror surface is less than 0.2.

The mirror distortion spectrum of Equations (24) to (27), subject to the assumptions of Equations (28) to (32), yields a roughness spectral density that falls off so rapidly with increasing spatial frequency f that the surface is essentially perfect on an optical scale, i.e., has practically no roughness at optical dimensions (due to scratches, grinding, and polish). To see this, the rms surface roughness of a small circular portion of mirror of diameter d will be computed. Measuring about the average surface distortion of this portion effectively throws away low-frequency distortion components (with wavelengths longer than d). Denoting this rms roughness by $m_{(d)}$ and substituting the value of Equation (30) into Equation (26) produce

$$m_{(d)}^2 \approx 2\pi \int_{1/d}^{\infty} \underset{\sim}{P}_m(f) f df = 2\pi \int_{1/d}^{\infty} M^2 \ \frac{\pi}{c^2} e^{-\left(\frac{\pi f}{c}\right)^2} f df$$

$$= \tfrac{1}{4} \times 10^{-8} \int_{(\pi/cd)^2}^{\infty} e^{-y} dy = \tfrac{1}{4} \times 10^{-8} e^{-\left(\frac{\pi}{cd}\right)^2} \qquad (33)$$

Thus, for a portion of mirror 100 microns $= 10^{-4}$ meter in diameter, taking the minimum value of the correlation length $1/c$ [Equation (32)] yields

$$m^2_{(10^{-4})} < \tfrac{1}{4} \times 10^{-8} e^{-(10\pi)^2} = \tfrac{1}{4} \times 10^{-436}$$

$$(34)$$

$$m_{(10^{-4})} < \tfrac{1}{2} \times 10^{-218} \text{ meter} = \tfrac{1}{2} \times 10^{-212} \text{ micron}$$

A similar calculation for a circular portion of the mirror 1 mm in diameter yields

$$m_{(10^{-3})} < 0.36 \text{ micron} \qquad (35)$$

These results decrease rapidly as the correlation length $1/c$ increases beyond its minimum value of 1 mm. The following table shows the diameter of a section of mirror which will be essentially diffraction-limited at $\lambda = 1$ micron, for a few correlation lengths:

M = 0.05 mm, rms total surface roughness

$m_{(d)}$ = 0.05 micron

$\frac{1}{c}$	0.1 cm	1 cm	10 cm
d	0.085 cm	0.85 cm	8.5 cm

On intuitive grounds alone, it can be assumed that the correlation length $1/c$ is of the order of a few centimeters. As noted above, the value of $1/c$ [and, more fundamentally, whether the Gaussian covariance and spectrum of Equations (24) to (26) are even appropriate] depend on the details of construction and manufacture, which are not considered here. The correlation length $1/c$ is consequently regarded as a parameter in the final results, within the range specified above, without further justification. Correlation functions of the form of Equation (24) have been used in the theory of random antennas[80] and in the theory of propagation through a random medium.[81]

While the above model provides one interesting case, the great surface accuracy assumed at optical dimensions makes it appear somewhat unrealistic. Therefore, the following alternative model will be considered.

Exponential Covariance. As a second example, assume an exponential covariance

$$\underset{\sim}{\wp}_m(\tau) = M^2\, e^{-c|\tau|} \tag{36}$$

with corresponding spectral density

$$\underset{\sim}{P}_m(f) = M^2 \frac{2\pi/c^2}{\left[1 + \left(\frac{2\pi f}{c}\right)^2\right]^{3/2}} \tag{37}$$

This form of covariance has also been used in the theory of propagation through random media.[78] M and c again have the physical interpretation of Equation (25). The total spectral content in the vicinity of spatial frequency f [here defined as $\underset{\sim}{P}_m(f)$ integrated over the elemental area $2\pi f df$] falls off as $1/f^2$ for the spectrum of Equation (37) [see the first relation of Equation (33)]. Thus the rms surface deviation is, in this sense, proportional to the (spatial) wavelength of the corresponding component of mirror distortion; this might appear physically plausible. The high spatial-frequency content of Equation (37) is clearly far greater than that of Equation (26). The rms gradient in this case is infinite, since the random surfaces are non-differentiable.[78] Making similar assumptions to those for Gaussian covariance [following Equation (27)], the conditions of Equations (29) to (32) are again required to apply in the present case (exponential covariance). Equation (32) insures that truncating $P_m(f)$ of Equation (37) at $f = 1/\lambda$ for $\lambda = 1$ mm $= 10^{-3}$ meter will reduce the mean square mirror distortion by less than 1.57 percent, corresponding to the minimum correlation length $1/c = 1$ mm. This number again decreases as $1/c$ increases and is much smaller at optical frequencies, $\lambda = 1$ micron $= 10^{-6}$ meter. However, truncating Equation (37) reduces the mean square gradient M' from ∞ to some finite value; thus the cusp at the origin in the covariance of Equation (36) is removed by the operation of Equation (23). Further study is necessary to determine if the modification in behavior for small τ may or may not be neglected in subsequent work.

The mirror distortion for exponential covariance [Equations (36) and (37) and the condition of Equations (29) to (31)] has significant roughness at optical dimensions, in contrast to the case of Gaussian covariance considered above. To illustrate this as in the Gaussian case above, Equation (37) is used to compute the rms surface roughness of a small circular portion of diameter d, measured about the average surface distortion of the portion:

$$m_{(d)} = \frac{M}{\left[1 + \left(\frac{2\pi}{cd}\right)^2\right]^{1/4}} = \frac{\tfrac{1}{2} \times 10^{-4}\cdot}{\left[1 + \left(\frac{2\pi}{cd}\right)^2\right]^{1/4}} \tag{38}$$

whereas in Equation (33) the value of M given by Equation (30) was used. By the restriction of Equation (19), $c < 10^3$ meters^{-1}, considering only circular portions of diameter $d < 2$ mm, then $(2\pi/cd)^2 > 9.86$, and Equation (38) may be approximated by

$$m_{(d)} \approx \tfrac{1}{2} \times 10^{-4} \sqrt{\frac{cd}{2\pi}}\ ,\quad \left(\frac{2\pi}{cd}\right)^2 \gg 1 \tag{39}$$

Consider circular portions of mirror 10 microns in diameter. For a correlation length $1/c = 1$ mm,

$$m_{(10^{-5})} = 2.0 \text{ microns} \tag{40}$$

for $1/c = 1$ m,

$$m_{(10^{-5})} = 0.63 \text{ micron} \tag{41}$$

The mirror tolerance at optical dimensions is clearly significant here, and may be varied from quite poor to quite good by changing the correlation length $1/c$ over the range of 1 mm to several cm. Again on intuitive grounds alone it might be assumed that the correlation length $1/c$ lies in the range of a few centimeters. However, it will appear below that the model of Equations (36) to (37) predicts optical behavior far worse than indicated by common experience for a well-polished reflector, e.g., a searchlight reflector. Rather, this model seems appropriate for a poorly polished aluminum reflector that is good for millimeter waves, but is so rough optically that it yields a very diffuse spot at optical frequencies.

5.2.1.3 Focal-Plane Field -- Gaussian Covariance. The mean-square value of the (unnormalized) focal-plane field for an incident field consisting of an ideal plane wave of unit intensity is[80]

$$<|w|^2> = \frac{1}{(\lambda F)^2}\, \mathcal{R}\left(\frac{r}{\lambda F}\right) \tag{42}$$

where r is the radius from the axis measured in the focal plane, λ the free-space wavelength, and F the focal length of the mirror.

The function $\mathcal{R}$ is given for the millimeter region (small rms phase error) by[78]

$$\mathcal{R}(f) = e^{-\left(\frac{4\pi M}{\lambda}\right)^2} A^2(f) + \left(\frac{4\pi^2 aM}{\lambda c}\right)^2 e^{-\left(\frac{\pi f}{c}\right)^2}$$

$$\left(\frac{4\pi M}{\lambda}\right)^2 \ll 1 \quad , \quad \lambda < \frac{1}{c} \ll 2a \tag{43}$$

In Equation (43), A(f) is the Airy disc function,

$$A(f) \equiv \pi a^2 \; \frac{J_1(2\pi fa)}{\pi fa} \tag{44}$$

The first term of Equation (43) corresponds to the average field in the focal plane. For a perfect mirror, M = O, the second term goes to zero and the first term represents the power in the usual diffraction pattern at the focal plane for an incident plane wave. As the mirror becomes worse and M increases, or alternately as the frequency increases and λ decreases, the contribution of the first term decreases and that of the second term increases; however, the approximation of Equation (43) fails before a very large fraction of power resides in the second term. Thus a different approximation is required for the optical region.

For the optical region (large phase error), the function $\mathcal{R}$ is given by

$$\mathcal{R}(f) = e^{-\left(\frac{4\pi M}{\lambda}\right)^2} A^2(f) + \left(\frac{\lambda a}{4cM}\right)^2 e^{-\left(\frac{\lambda f}{4cM}\right)^2}$$

$$\left(\frac{4\pi M}{\lambda}\right)^2 \gg 1 \;,\; 4M \ll \frac{1}{c} \ll \left(\frac{4\pi M}{\lambda}\right) 2a \tag{45}$$

The first term of Equation (45) is identical to the first term of Equation (43); however, here the first condition of Equation (45) guarantees that the first term, which represents the average focal-plane field, may be neglected. The second condition guarantees that the small-angle approximations, implicit in the above results, remain valid.

Equation (45) may be written in an alternative way that is illuminating. Neglecting the first term of Equation (45) and using Equations (42) and (27),

$$<|w|^2> = \left(\frac{a}{4cMF}\right)^2 e^{-\left(\frac{r}{4cMF}\right)^2} = \left(\frac{a}{2M'F}\right)^2 e^{-\left(\frac{r}{2M'F}\right)^2} \tag{46}$$

gives the expected (unnormalized) intensity at the focal plane, subject to the conditions in Equation (45). Equation (46) does not depend on λ (or, equivalently, on the frequency) of the incident plane wave. Equation (46) is easily derived from geometric optics; the only relevant statistical property of the random surface is its rms gradient.

Focal-plane field - exponential covariance. For the millimeter-wave region (small phase error), and using Equations (22) and (37) above,

$$\mathcal{R}(f) = e^{-\left(\frac{4\pi M}{\lambda}\right)^2} A^2(f) + 2\left(\frac{4\pi^2 aM}{\lambda c}\right)^2 \frac{1}{\left[1 + \left(\frac{2\pi f}{c}\right)^2\right]^{3/2}}$$

$$\left(\frac{4\pi M}{\lambda}\right)^2 \ll 1 \;,\; \lambda < \frac{1}{c} \ll 2a \tag{47}$$

The interpretation of Equation (47) is similar to that given above for Equation (43); their first terms are identical. Equation (47) decreases much more slowly for large f (corresponding to the tails of the focal-plane spot) than does Equation (43).

In contrast to the previous case (Gaussian covariance), the convolution of Equation (23) – which corresponds to the truncation of the spectral density in Equation (22) – modifies the behavior of $\underset{\sim}{\phi}_{\gamma}(\tau)$ near the origin, from cusp-like behavior to parabolic behavior.

In the optical region (large phase error),

$$\mathcal{R}(f) = e^{-\left(\frac{4\pi M}{\lambda}\right)^2} A^2(f) + \left(\frac{\lambda a}{2M'}\right)^2 e^{-\left(\frac{\lambda f}{2M'}\right)^2}$$

$$\left(\frac{4\pi M}{\lambda}\right)^2 \gg 1 \;,\; 50\,\frac{M^2}{\lambda} < \frac{1}{c} \ll \left(\frac{4\pi M}{\lambda}\right)^2$$

$$\frac{2\pi a^2}{\lambda} \;,\; \frac{1}{c} < \left(\frac{4\pi M}{\lambda}\right)^2 (.2\lambda) \tag{48}$$

where

$$M' = M\sqrt{\frac{2\pi c}{\lambda}} \qquad (M' \gg M) \tag{49}$$

and A(f) is again given by Equation (44). The first term is negligible, and the second yields, via Equation (42), the identical result to the righthand member of Equation (46); this is again the geometric-optics approximation.

The first condition in Equation (48) again requires large phase errors; the second combines the small-angle approximation and the δ-function approximation in evaluating the convolution of Equation (35) of Reference 78, and the final condition is associated with the parabolic approximation.

For M given by Equation (30) and for λ = 1 micron, the correlation length must be less than 8 cm for the parabolic approximation to hc'd. while the same quantity must be greater than 12.5 cm for the small angle approximations to be valid. While strictly speaking there is no overlapping region, assume for

$$\frac{1}{c} = 0.1 \text{ meters} = 10\,\text{cm} \qquad (50)$$

that Equation (48) gives a useful approximation. For smaller correlation lengths 1/c the mirror becomes so poor that the small-angle approximations fail (and the mirror ceases to be of interest). Much larger correlation lengths do not appear to be of interest here, on intuitive grounds; if they were, the problem could be treated but the parabolic approximation would fail.

5.2.1.4 Numerical Example – Gaussian Covariance

$\mathcal{L}_m(\tau) = M^2 e^{-(c\tau)^2}$ [Equation (24)]

M = 0.5 x 10^{-4} meters, rms total surface roughness [Equation (30)]

2a = 10 meters, mirror diameter

F = 10 meters, focal length (i.e., assume an f/1 mirror)

Focal-Plane Field at λ = 1 mm = 10^{-3} Meters

$$<|w|^2> = \frac{1}{(10^{-3}F)^2} \left\{ 0.674 A^2 \; \frac{r}{10^{-3}F} + 3.89 \left(\frac{a}{c}\right)^2 e^{-\left(\frac{\pi}{c} \; \frac{r}{10^{-3}F}\right)^2} \right\}$$

$$10^{-3} < \frac{1}{c} \ll 2a = 10 \text{ meters} \qquad (51)$$

The first term of Equation (51) is the diffraction pattern of an ideal antenna [Equation (47)], reduced in total power to 67 percent by the imperfections; the second term represents the power scattered by the imperfections, here 39 percent of the total. (These two numbers do not add up to 100 percent because of the inaccuracy in the small phase error approximation. The rms phase error here is $\Gamma = 2\pi \times 1/10$ radians = 36 degrees, which is significant compared to 2π radians = 360 degrees.) The angular half-widths of the focal-plane spots of the two components are roughly:

$\left(\frac{r}{F}\right)_{av.} = 1.2 \times 10^{-4}$ radian [average component – first term of Equation (51)]

(52)*

In view of the restriction on c [Equation (51)],

$$10^{-4} \ll \left(\frac{r}{F}\right)_{random} < 1 \qquad (53)$$

For sufficiently small correlation length, 1/c < 5.3 meters, the spot size of the random component will be larger than that of the average (or coherent) component; the previous intuitive idea that 1/c ~ a few centimeters thus leads to a random component several orders of magnitude larger in diameter than the coherent component. In particular, for F = 10 meters, the spot sizes become:

$\frac{1}{c}$ (cm)	0.1	1	10	
r_{av} (mm)	1.2	1.2	1.2	
r_{random}(m)	6.36	0.636	0.0636	(54)

Focal-Plane Field at λ = 1 Micron = 10^{-6} Meter

$$\langle|w|^2\rangle = \frac{1}{(10^{-6}F)^2} \; 0.25 \times 10^{-4} \left(\frac{a}{c}\right)^2 e^{-\left(\frac{1}{200c} \; \frac{r}{10^{-6}F}\right)^2}$$

$$= 2 \times 10^{-4} \ll \frac{1}{c} \ll 6.28 \times 10^{3} \text{ meter} \qquad (55)$$

The angular half-width of the focal-plane spot is approximately

$$\frac{r}{F} \qquad \frac{r}{F} = 4 \times 10^{-4}\, c \qquad (56)$$

From the restriction on c [Equation (55)],

$$0.636 \times 10^{-7} \ll \frac{r}{F} \ll 2 \qquad (57)$$

For F = 10 meters as assumed above,

$\frac{1}{c}$ (cm)	0.1	1	10	
r (cm)	400	40	4	(58)

r is approximately the required radius of the optical detector.

5.2.1.5 Numerical Example – Exponential Covariance

$\mathcal{L}_m(\tau) = M^2\, e^{-c(\tau)}$ [Equation (36)]

M = 0.5 x 10^{-4} meter, rms total surface roughness [Equation (30)]

2a = 10 meters, mirror diameter

F = 10 meters, focal length (i.e., assume an f/1 mirror).

*Spot radius of the average component is defined quite arbitrarily at the first zero of the Airy function A(f) [Equation (45)]. Spot radius of the random component is defined such that 98 percent of the total power is included. For Gaussian spots, as in the present case, halving the radius will reduce the power to 63 percent of the total.

Focal-Plane Field at $\lambda = 1$ mm $= 10^{-3}$ Meter.

$$\langle |w|^2 \rangle = \frac{1}{(10^{-3}F)^2} \left\{ 0.674\, A^2 \left(\frac{r}{10^{-3}F} \right) + 3.89 \left(\frac{a}{c} \right)^2 e^{-\left(\frac{\pi}{c} \frac{r}{10^{-3}F} \right)^2} \right\}$$

$$10^{-3} < \frac{1}{c} \ll 2a = 10 \text{ meters} \qquad (59)$$

The discussion following Equation (51) for Gaussian covariance, up to and including Equation (57), applies to the present case of exponential covariance also, with the exception of the spot size for the random component, $(\frac{r}{F})_{\text{random}}$. The distribution of power in the random component differs in the two cases, falling off much more slowly in the tails of the focal-plane spot for exponential covariance [Equation (59)] than for Gaussian covariance [Equation (51)]. For 98 percent of the total power included in the spot, the random component for exponential covariance is 12.5 times larger than that for Gaussian covariance [Equation (52)], for the same correlation length 1/c; however, for 63 percent of the total power included, the random component for the present case is only 1.25 times larger.

Focal-Plane Field at $\lambda = 1$ Micron $= 10^{-6}$ Meter. Equation (50) restricts the discussion to a single correlation length, 1/c = 10 cm. From Equation (49)

$$M' = 0.396 \text{ for } \frac{1}{c} = 0.1 \text{ meter} \qquad (60)$$

Then

$$\langle |w|^2 \rangle = \left(\frac{a}{0.792F} \right)^2 e^{-\left(\frac{r}{0.792F} \right)^2}, \quad \frac{1}{c} = 0.1 \text{ meter} \qquad (61)$$

The angular half-width of the focal-plane spot is approximately

$$\frac{r}{F} = 1.58, \frac{1}{c} = 0.1 \text{ meter} \qquad (62)$$

Thus for F = 10 meters,

$$r = 15.8 \text{ meters} \qquad (63)$$

where Equations (61) and (62) have been computed on the same basis as before. Clearly this mirror is very poor; further, as the correlation length 1/c decreases it will get much poorer still (and the analysis will fail completely).

5.2.1.6 Discussion. It appears that the use of a reflector good to millimeter-wave tolerances in an incoherent optical receiver will require detectors of substantial diameter (or alternatively large arrays of small detectors and associated amplifiers). The example of Section 5.2.1.5 is optimistic in two respects:

1. The Gaussian covariance predicts a surface much too smooth at optical dimensions.
2. An f/1 mirror has been assumed. (Doubling the focal length doubles the spot size and hence the required detector diameter.)

The requirement that the antenna be good at $\lambda = 1$ mm (i.e., suffer a decrease in gain of 1.7 dB at $\lambda = 1$ mm) imposes a roughness tolerance of about 2 mils rms. The correlation length 1/c of the surface roughness is crucial; the larger it is the better. The table of Equation (58) shows that for 1/c = 10 cm a detector 8 cm in diameter is required (to receive 98 percent of the incident power), while for 1/c = 1 cm a detector 80 cm in diameter is required.

Thus, to achieve a detector diameter of 8 cm ~ 3 inches with a 33-foot diameter f/1 mirror, its surface tolerance must be 2 mils, the surface distortion correlation distance exceeding 4 inches. This would seem to imply a large area vacuum photomultiplier as a detector, with the consequent restriction to the visible portion of the spectrum.

The above results imply a large acceptance angle for background radiation in such a system. In the example discussed in the preceding paragraph [Section 5.2.1.5, Equation (58), 1/c = 10 cm], the half-angle defining the field of view of the receiver is 4×10^{-3} radians, corresponding to a solid angle of $16\pi \times 10^{-6} \approx 5 \times 10^{-5}$ steradian.* For the shorter correlation lengths in Equation (58), these angles will be correspondingly larger, with poorer noise performance resulting (see Section 5.7).

The above results (see also Appendix 3) indicate that optical collectors can not be built with millimeter tolerances unless the correlation length is several meters. Stated somewhat differently, when surface errors are large compared to a wavelength, then the angular field of view is determined essentially by the rms deviation of the slope of the surface. Assuming correlation lengths of the order of 10 cm, then the surface errors would have to be less than 2 microns to achieve a field of view of the order of 10^{-4} radian.

5.2.2 Possibilities for an Incoherent Optical Receiver

To put the above results otherwise, average surface slopes must be so small that the surface, when viewed by eye, has a characteristic optical polish and will look like a good mirror. This suggests possibilities for an incoherent receiver which may be more fruitful then the coated millimeter-antenna approach. In particular, solar furnaces

*The effective solid angle to background radiation is given by $\Omega = \pi r^2/F^2$, independently of the reflector imperfection statistics, where r is the radius of the photodetector.

answer rather well the description of an incoherent optical receiver. They are designed to form an image of the sun with sufficiently small error that the image is not significantly larger than that obtained with a perfect reflector (i.e., one having negligible surface errors). Of course, the sun subtends 32 minutes of arc at the earth, so this requirement is rather loose for present purposes. Nevertheless, these are devices which have been built and successfully operated, and may offer some realistic guidelines.

It is not useful to survey solar furnaces broadly – only to examine the newest, largest, and best of them. The solar furnace operated by the Army Quartermaster Corps at Natick, Massachusetts appears to be the largest in the United States. This has a stationary spherical "concentrator" and a movable planar reflector (or heliostat) in order that the work at the focus can be stationary. The concentrator comprises 180 two-foot-square concave mirrors in a square array which is 14 mirrors on each edge. (There is a blind spot in the center, so there is a 4 mirror by 4 mirror hole). The useful area is 720 square feet and the focal length is 35 feet. According to Fred Penniman,[82] each mirror, made by the American Optical Company, is 1/4 inch thick and cost $70. Thus the total mirror cost was $12,600. The frame cost was about $50,000, studding was $10,000, and alignment cost $2,000.[82] Total reflector cost was therefore about $75,000. This is in agreement with the general approximate rule given by Daniels[83] that solar furnaces with high optical precision may cost over $100 per square foot. The resolution of the reflector was found by E. Fazio to be about 1/4 degree.[82] Even at that, thermal expansion of the supporting structure (which has no enclosure) causes loss of resolution. The best mirror surface is aluminum, overcoated with silicon monoxide, but even this does not last more than a year. The mirrors are regularly recoated and replaced.

Perhaps the most interesting solar furnace is at Tohoku University, Sendai, Japan.[84] This is a paraboloid with a 10-meter-diameter circular aperture. The reflector is set back into a building, and doors enclose it when it is not in use. The reflector is segmented into seven concentric bands, each containing a number of mirror panels. The mirrors are front-aluminized by vacuum-evaporation, but no silicon monoxide overcoat is used, and the mirrors must be recoated regularly. The panels were prepared individually, each shaped to its part of the paraboloid, although all panels in one band have the same shape. The reflector is remarkable in that the tolerance for angular error in the reflected rays is only 1/2 milliradian. This furnace thus demonstrates that an incoherent receiver can be built, at reasonable cost, having a field or view of 1/2 milliradian. Actual cost was not stated,[84] and an inquiry regarding this has not yet been answered.

Beyond solar furnaces, there is little definite information regarding technical feasibility of incoherent receivers except for two studies. Sylvania[85] is studying the design of a paraboloid with a 10-meter-diameter aperture and a 10-meter focal length. This would have about 800 molded epoxy mirrors each with an area of about 0.1 square meter. The mirrors would be servo-positioned to 1/4 wave length of visible light, and a field of view of 0.1 milliradian would be obtained. The estimated cost of each mirror is about $30. This is quite inexpensive considering the $70 figure for the mirror cost of the Natick solar furnace and considering that the mirrors for the receiver would be optically much superior to those in the furnace. Sylvania's estimate for total cost of the receiver is $1 million.

Perkin-Elmer has reported[86] a study of incoherent receivers, giving some technical detail but no cost estimates. (No study of cost as a function of size or tolerance seems to have beem made by anyone.) Perkin-Elmer considers a spherical primary mirror comprising many segments, perhaps hexagonal, and all made by replication for simplicity and low cost. A secondary package would rotate about the center of curvature of the primary. This would contain a secondary and a tertiary mirror for all-reflecting optics (suitable for use at 10.6 microns). The uncorrected primary would display such spherical aberration that the field of view would be quite large. The secondary and tertiary mirrors would correct the spherical aberration to a circle of confusion of 6 arc seconds. The mirrors also provide a large effective focal length, so that the f-number would be f/30. This is sufficiently collimated so as not to increase the passband of narrowband interference filters. The use of all-reflecting optics leads to a design in which the middle third, or a little more, of the primary diameter used is obscured by the limited acceptance of the secondary and tertiary, although this is not inconsistent with the precedent of the loss (or blockage) in Cassegrain-type feeds.

A field stop would provide a 20 arc second field of view. This would be about the smallest field which would be attainable by such a structure when all degradations are included. If each segment had a 40 inch diameter and there were a 0.0004 inch tilt on that diameter, the reflected rays would have an angular error of 4 seconds. If the tilts were randomly oriented, the initial 6 arc second circle of confusion would grow to an overall diameter of 14 arc seconds. When atmospheric seeing is added (especially the relatively poor, but unknown, daytime seeing), the total field of view would come to 20 arc seconds (100 microradians) at least.

The tolerance on mirror tilt suggested here is only one part in 10^5. This is small enough that considerable attention would have to be paid to thermal distortions. (Recall the sensitivity even of the Natick solar furnace with its 1/4 degree field of view.) With ordinary materials of construction, temperature differences of the order of a degree would misalign this receiver. A supporting structure of invar would be technically attractive, but, on the scale of this structure, economically prohibitive. The receiver would certainly have to be enclosed, and temperature uniformity would have to be attained through rapid circulation of conditioned air. The cost of such an enclosure would be a major part of the total cost, and it would increase very rapidly with the size of the receiver. This tolerance on mirror tilt also would require regular realignment of the mirrors. If active control of the kind proposed by Sylvania were not used, then at least the segments would require periodic (perhaps nightly) testing and adjustment.

Two types of primary structure were considered by Perkin-Elmer. The primary would be very large and heavy, and, to avoid turning it, one possibility is to make it a fixed hemispherical bowl. For an aperture of 10 meters, as seen by the secondary package at any one time, the bowl would have a diameter of perhaps 30 meters or more. Hence, to avoid turning the primary, a huge structure results in which only a few percent of the total reflector would be included in the field of the secondary at any instant. The enclosure would be correspondingly large and expensive. The secondary would be in an elevation-azimuth bearing, of the type described previously for the diffraction-limited receiver, except now turned upside down, with the azimuth bearing hanging from a frame suspended over the bowl.

An alternative would be a primary whose shape would be only part of a spherical zone 10 meters wide. This would stand vertically on a very large azimuth bearing. Here the disadvantage would be the large mass the azimuth bearing must carry, but the great advantage is a large reduction in necessary size of structure and enclosure. The secondary package here would turn on an elevation bearing supported by the same azimuth bearing.

It is impossible now to choose between these (and perhaps other) alternatives for the type of primary structure. This would require an extensive design and tradeoff study. However, in view of the scale of total structure required for a fixed primary, one would hope that careful and resourceful design of a large azimuth bearing would permit a steerable primary at far lower cost than that for a fixed primary.

The Perkin-Elmer study gave considerable weight to the need for sky coverage down to the horizon and to the resulting difficulties of design. The problems of optical attenuation, refraction, and noise militate strongly against small elevation angles. Since an array of stations around the world would be required in any case, it would appear likely that lowest overall cost would result from coverage at each station not below about 30, or perhaps 20, degrees.

6. OPTICAL FILTERS

The main purpose of a predetection optical filter is to increase the signal-to-noise ratio of the radiation falling on the detector. This is achieved by reducing the background noise that is present outside the signal band by using a narrowband filter and at the same time minimizing self-emission noise from the filter itself. Since the ~10^6 to 10^7 Hz signal band of a deep space optical communication system is very small relative to the filter passband, the only filters of interest in the present study are narrow bandpass types.

Optical filters with narrow bandpass are based on the interference principle, of which there are three basic types. They are thin film filters, Fabry-Perot filters, and birefringent filters. The first two employ interference of direct and multiple reflected wave components. The third is based on interference of differentially retarded orthogonal polarization components through a birefringent medium. A brief description of each type is given in Section 6.1. The next two sections consider the state of the art of these filters for visible and infrared regions. Recommendations are made in Section 6.4.

6.1 General Principles

6.1.1 Thin-Film Interference Filter

Filters of this type consist of a set of many alternate quarter-wave layers of high and low refractive index dielectric films that are separated from a similar set by a half-wave dielectric spacer. Since the dielectric films have low loss, high transmission can be achieved at the peak of the passband. If T_1 and T_2 are respectively the transmissions of the two sets of dielectric reflectors, and R_1 and R_2 their respective reflectances at the interface of the spacer, then the peak transmission, T_{max}, at the center of the band, is given by[87]

$$T_{max} = \frac{T_1 T_2}{(1 - R)^2} \qquad (64)$$

where $R^2 = R_1 R_2$. It can be observed from Equation (64) that, for nonabsorbing and nonscattering dielectric films, a transmission of 100 percent can be achieved. The half-power bandwidth is given by[87]

$$\Delta\lambda = \frac{2\lambda}{m\pi - \frac{d\Phi}{d\lambda}} \arcsin \frac{1 - R}{2\sqrt{R}} \qquad (65)$$

where m is the order of interference, λ the wavelength of operation, and $d\Phi/d\lambda$ is the average rate of phase change upon reflection from the two reflecting systems. Typically, $d\Phi/d\lambda$ is assumed to vary linearly over the width of the passband. It can be observed from Equation (65) that in principle the bandwidth can be made arbitrarily small by increasing R toward unity, which in turn can be achieved by increasing the number of layers in each set of reflectors.

However, due to technological difficulties involved in achieving uniformity and evenness of multilayer coatings, the bandwidth attained in practice is limited to small but finit values, as will be discussed in more detail later in this section.

The thin-film interference filter technique has been used in visible, near infrared, and far infrared wavelength regions. Thin-film technology is well advanced and this filter type has advantages of large aperture size, wide field of view, and good temperature stability. It is relatively insensitive to small mechanical vibrations and normally needs no supplementary blocking filter. However, it suffers from the disadvantage that the minimum attainable passband, as limited by technology, is wider than that for the other two filter types.

A slight tilt of the filter from normal incidence shifts the passband to a shorter wavelength. Thus, orientation of the filter at a slightly tilted angle in the neutral position provides tunability, by slight perturbations about this position, to higher or lower wavelengths.

6.1.2 Fabry-Perot Filter

Basic construction of the Fabry-Perot filter is the same as that of the thin-film type except that the spacer layer is a thick, self-supporting structure. Spacing between the two sets of coatings is equal to an integral number of half-wavelengths. Thus, Equations (64) and (65) apply to the present case also. Because the half-wave separator plate is of higher quality in the Fabry-Perot structure, the dielectric reflecting layers can be coated more evenly. Thus, bandwidths can be achieved which are narrower than in the case of the thin-film type. Furthermore, it can be observed from Equation (65) that the width of the filter passband can be reduced by increasing the interference order m. This is achieved in the Fabry-Perot filter with a thicker spacer. However, since adjacent resonant modes (or passbands) are separated by $\Delta\lambda = \lambda/m$ in a Fabry-Perot cavity, an increase in m decreases the separation between adjacent modes. Consequently, a blocking filter generally must be used with Fabry-Perot filters to mask the unwanted passbands.

Very narrow bandwidth filters have been achieved at visible and near infrared wavelengths using the Fabry-Perot technique. This can easily be extended to the 10.6μ wavelength region, though no reports of development of such filters could be found in the literature. These filters have many advantages over the thin-film interference filter. The passband is narrower and the transmission higher. It can be tuned over a wide range, which is limited primarily by the blocking filter characteristics, by varying the angle of incidence. Continuity in wavelength range is achieved by using adjacent higher or lower order modes. Performance of the Fabry-Perot filter compares well with the thin-film interference filter with respect to temperature stability, aperture area, and field of view. One major drawback of the Fabry-Perot filter is that a blocking filter of a few tens of Angstroms bandwidth (e.g., $\Delta\lambda = \lambda/m = 20\text{Å}$ when $\lambda = 5000\text{Å}$ and $m = 250$) must be used.

6.1.3 Birefringent Filter

The birefringent filter, originally proposed by Lyot,[88] can be produced in many modified forms. These filters have been used in the past largely in connection with solar observation. They differ from the thin-film and Fabry-Perot filters in principle of operation and complexity of construction. They are designed on the principle of interference of orthogonal polarization components that have passed through layers of birefringent crystals. The direction of propagation is perpendicular to the plane containing the optic axes in the biaxial crystals and in any direction perpendicular to the optic axis in uniaxial crystals.

The birefringent filters can be designed to have very narrow bandwidths, though this may require very precise alignment of crystal axes. They are also tunable. The tuning is done by rotation, about the filter axis, of one or more of the several crystals arranged in series with respect to the others. The field of view of the filters is typically a few degrees. A procedure for synthesis of filters with desired passband characteristics has been proposed by Harris, Amman, and Chang[89] and has been investigated by Amman and Yarborough.[90] This method affords more versatility and a reduction in the number of components by a factor of two is claimed.

The birefringent crystals are usually long and are very sensitive to temperature changes. Besides, the incident polarization must be linear and aligned with a specific transverse axis of the birefringent filter. This alignment is insured with a polarizer element at the filter input, which in general introduces optical loss unless special modifications of the filter device are made to accommodate arbitrarily polarized incoming radiation. Such modifications, which have been proposed by Evans[91] and Amman,[92] reduce the effective aperture size. The birefringent filter also requires a blocking filter (of the same order of bandpass as for the Fabry-Perot type) to stop unwanted passbands. Consequently it suffers from the same disadvantage mentioned above in connection with the Fabry-Perot filter.

6.2 Filters for the Visible and Near-Infrared Regions

The performances of filters at visible and near infared wavelengths are comparable and hence have been grouped under a single heading. The optimum filter type for these regions depends on the specific mission. A compromise generally must be made between various parameters which include passband, transmission, angular sensitivity, detector area and allowed temperature fluctuations. The discussion here will be restricted to the present art for each filter type.

Of the three types, the thin-film filter has received the most developmental attention and probably has very closely approached the lower limit on the width of the bandpass. The primary problem is very small unevenness in the films. A bandwidth of about 1.5Å at 6328Å wavelength with ~30 percent transmission is attainable. Translating this to near infrared waveiengths, the same fractional bandwidth of about 0.025 percent can be expected. Temperature stability is very good (< 0.1Å shift in wavelength per °C and negligible change in bandwidth). The range of temperature that the filter can withstand mechanically is very wide (-160°C to 50°C is typical). The aperture diameter can be made fairly large – more than 2 inches; also the field of view is wide – about 5 degrees at normal incidence. A change in the angle of incidence of more than a few degrees shifts the band center noticeably toward shorter wavelengths, as mentioned previously. The magnitude of the wavelength shift Δλ for small angles of incidence can be theoretically calculated from the relationship[93]

$$| \Delta\lambda | = \frac{P\lambda_O \theta^2}{2} \tag{66}$$

where λ_O is the center of the passband at normal incidence, θ is the angle of incidence, and P is a constant that depends on the number of layers in the reflecting films and the order m of the spacer. There is, in general, fair agreement between the calculated and measured values. Figure 83 shows the experimental results for three dielectric thin film filters.[94] The percent shift to shorter wavelength is plotted as a function of angle of incidence, which, as may be noted from the figure, is relatively insensitive to the percent bandwidth. It has also been experimentally found that the bandwidth is relatively insensitive to the angle of incidence.[94]

The angular field of view for the thin-film filter varies with the angle of incidence. The angular field of view is here defined as the solid angle for which the magnitude of the wavelength shift with angle of incidence is less than or equal to 1/10 of the 3 dB passband. The transmission at these limits on the field of view should be 90 percent or more of the maximum transmission. A set of curves for the variation of field of view with incident angle for different bandwidths is given in Figure 84. The field of view initially increases with the angle of incidence, reaches a maximum, and then decreases to a nearly constant value. It appears advantageous to design the filter for a few degrees off normal incidence for two reasons. First, the filter can be tuned to lower or higher wavelengths by respectively increasing or decreasing the angle of incidence. Second, the change in the angular field of view with angle of incidence can be minimized. Figure 85 summarizes transmission vs. fractional bandpass Δλ|λ for the best commercially available thin-film filters. It may be observed from the figure that the transmission decreases as Δλ|λ approaches zero.

Fabry-Perot filters can provide a narrower bandwidth than the thin-film type. Herriot[95] has made a Fabry-Perot filter for 6328Å radiation with (1/2)Å bandwidth and 60 percent transmission. The separation between adjacent bands is 15Å, and consequently the device requires a blocking filter of the thin-film type. The angular dependence of Fabry-Perot filters is very similar to that of the thin-film filter. Consequently, the discussion given earlier in this section for the thin-film filter applies to the present case. The wavelength shift with angle of incidence also is toward shorter wavelengths; however, the Fabry-Perot filter is superior to the thin-film filter in that the wavelength can be increased by increasing the angle of incidence sufficiently to pick up the next lower order mode. However, this will not be a convenient arrangement to use in practice. Figure 86 gives the theoretical and experimental results on the wavelength shift with angle of incidence for a Fabry-Perot filter with a mica substrate as the spacer.[87] It would be advantageous, as in the case of the thin-film filter, to design the filter for a few degrees off-normal incidence so that the band center of a single mode can be shifted to lower or higher wavelength by respectively increasing or decreasing the angle of incidence.

The angular field of view for a Fabry-Perot filter is given in Figure 87 with width of passband as a parameter. As in the case of the thin-film filter, it is again better to operate at a few degrees off normal incidence for minimizing dependence of the angular field of view on changes in angle of incidence.

From the points of view of aperture area and filter thickness, the birefringence filter is the least attractive. The aperture area is restricted by size limitations of the birefringent crystal; minimum length of the crystal is dictated by the narrowness of the filter bandwidth and the birefringence characteristics of the crystal used. Steel et al [96] have built a birefringence filter of the Lyot type, a few inches in length, that has a bandwidth of (1/8)Å in the visible region and is tunable over ±16Å. As mentioned previously, the birefringent filter also needs a blocking filter. Transmission of the filter that Steel et al built is 12 percent, including that of the blocking filter. Another disadvantage of the bifringent filter is its extreme sensitivity to temperature. A temperature change of 0.01°C produces a change in the differential retardation of 1 degree in the above example. Consequently, temperature stability better than 0.01°C will be desirable. For simplicity of design, the temperature is maintained about 10°C above ambient by placing the filter in a temperature-controlled oven. The angular field if view is 2–1/4 degrees. The filter is tuned by rotating the relative crystal orientations about the axis of the filter. Consequently, tuning does not significantly alter the bandwidth or the field of view of the filter. A modified version of this filter type, the basic principle of which is described in Reference 89, is under development.[97] This has the added advantage over the conventional birefringent filter in that the shape of the passband can be tailored to specifications.

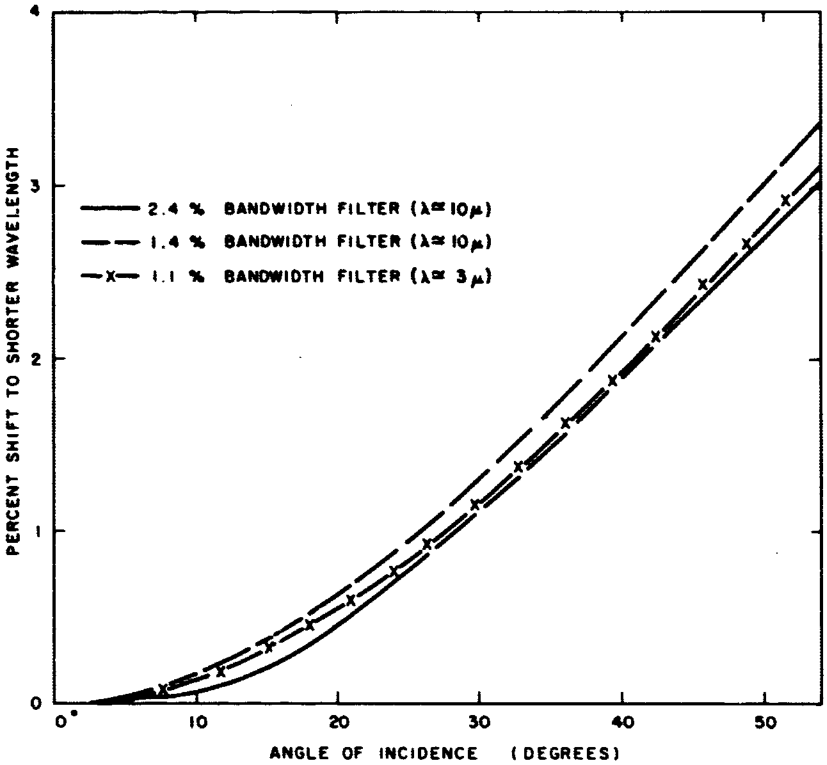


Figure 83. Wavelength shift of three thin-film narrowband filters vs. angle of incidence

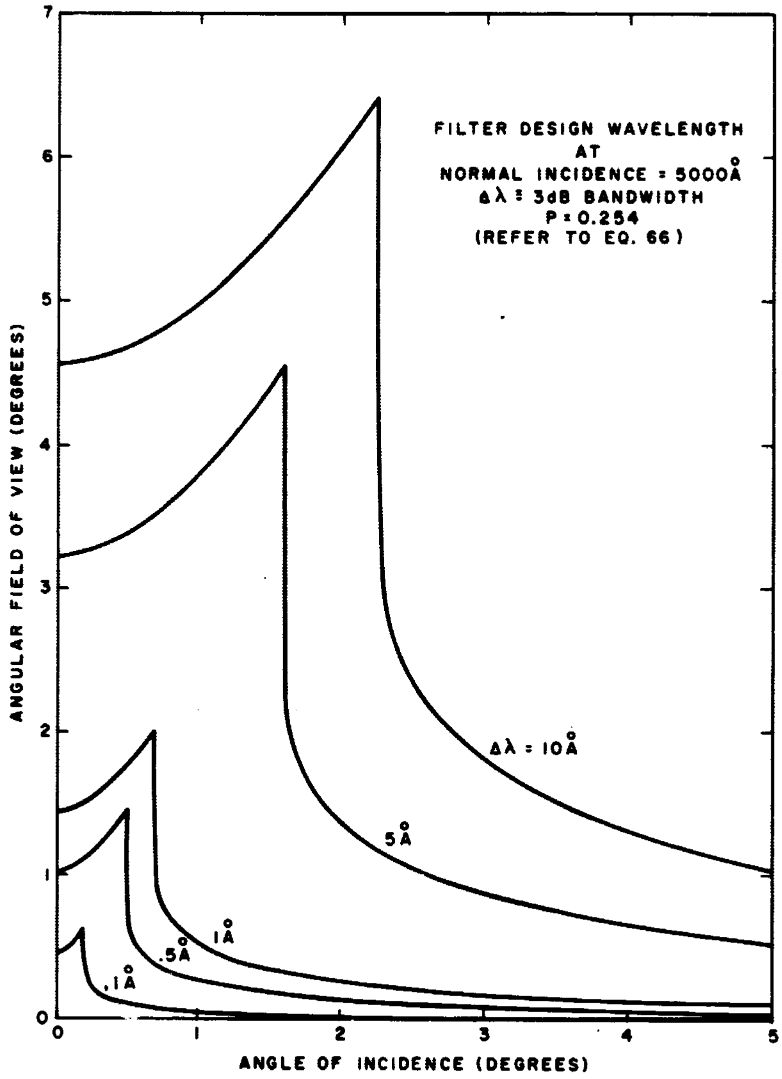


Figure 84. Field of view vs. angle of incidence for a thin-fil:n filter

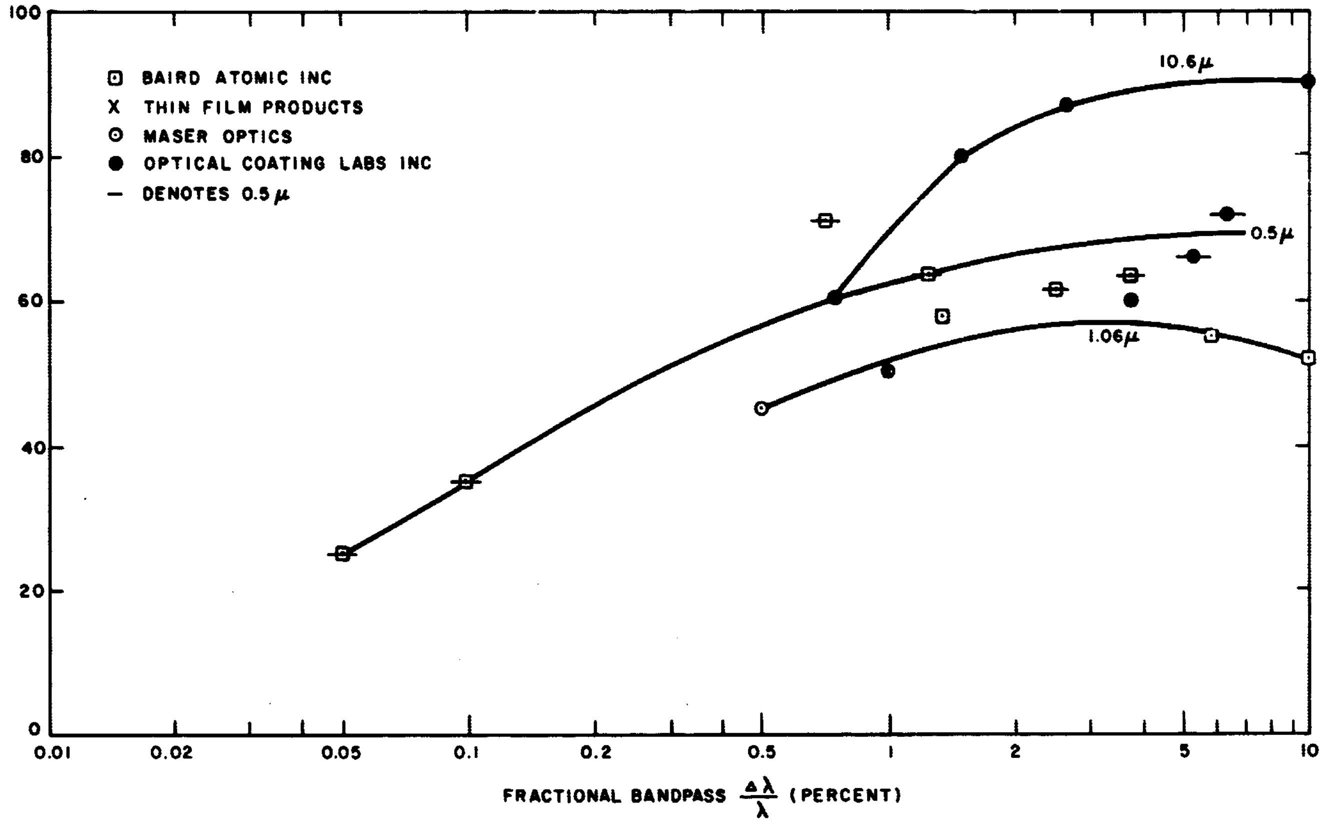


Figure 85. Peak transmission, T, vs. fractional bandpass Δλ/Δλ for commercially available thin-film filters at 0.5, 1.06, and 10.6 micron wavelengths

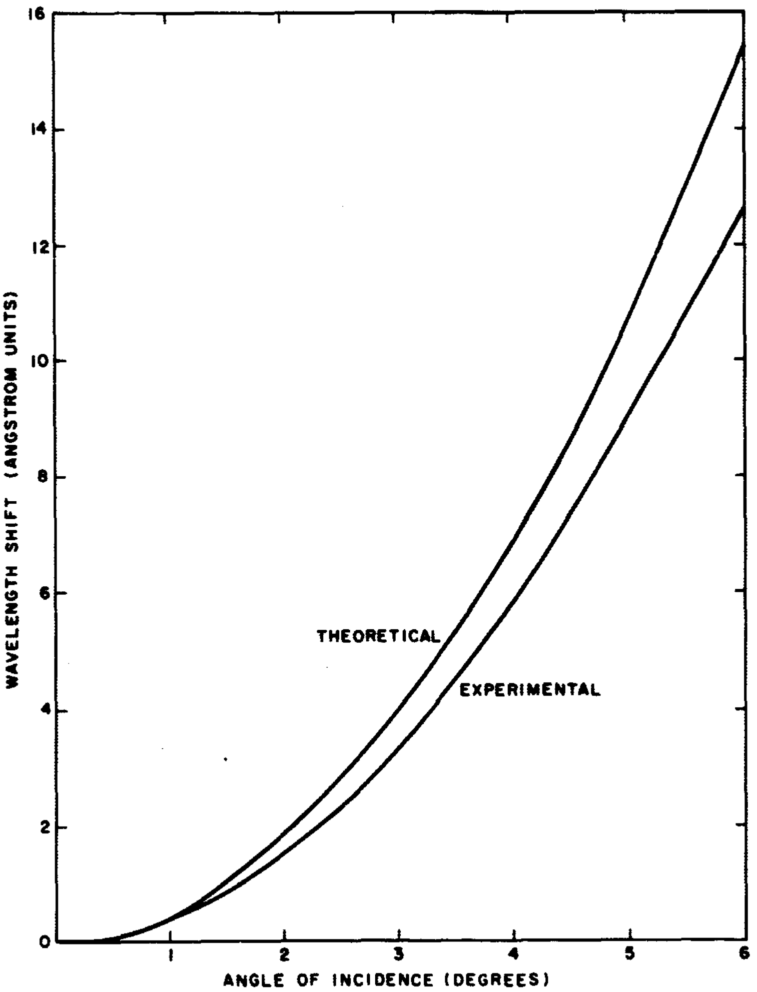


Figure 86. Wavelength shift vs. angle of incidence for a Fabry-Perot filter

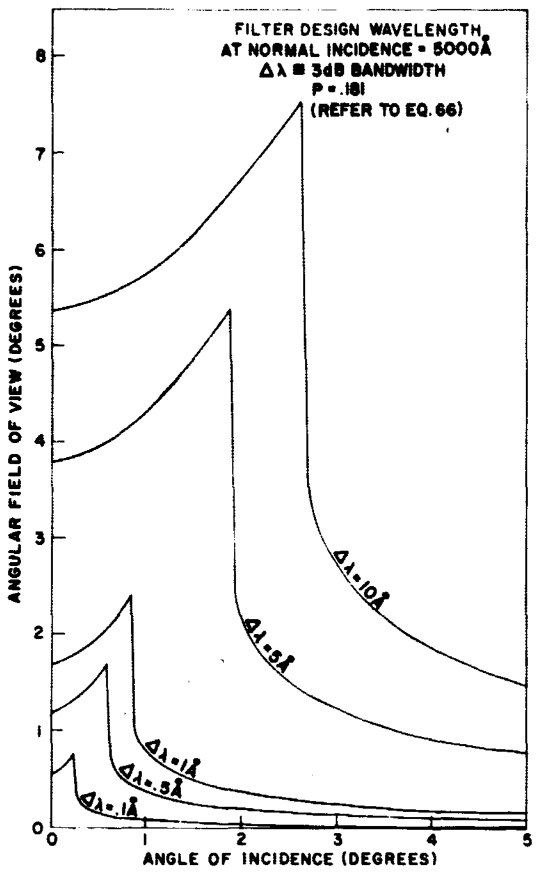


Figure 87. Field of view vs. angle of incidence for a Fabry-Perot filter

6.3 Filters for 10.6μ Wavelength Region

The only narrowband filter type on which published information could be found for this spectral region is the thin-film interference filter. Specifications on a commercially available filter[98] contain about 3/4 percent fractional bandwidth. This corresponds to about 760Å bandwidth at 10.6μ. Typical transmission is about 50 to 60 percent. It has about the same order of temperature stability as those described for visible region (Section 6.2). A tunable filter in this region[99] has been produced by deposition of a filter with continuously variable thickness on a circular disc. The angular field of view for commercially available 10.6μ filters for normal incidence is on the order of a few degrees. Figure 85 shows transmission and fractional bandwidth characteristics of the best commercially available thin-film filters at 10.6μ.

It is technologically feasible to produce a Fabry-Perot filter at 10.6μ wavelength, with Irtran as one possible substrate or spacer material. With this type, one can expect to achieve bandwidths in the range of a few tens of Angstroms. Due to lack of crystals with large birefringence at 10.6μ, it is not apparent that an efficient 10.6μ birefringent filter can be produced. No work on either Fabry-Perot or birefringent narrow-band filters at 10.6μ seems to have been reported.

6.4 Conclusions

For the visible and infrared wavelength regions, three types of narrow-band filters are available – thin-film, Fabry-Perot, and birefringent filters. The narrowness of the passband is limited to approximately 1.5Å in the thin-film filter, whereas the other two types can be made narrower (Table 33). The birefringent filters suffer from the major disadvantages of long lengths (several inches) and sensitivity to temperature change ($\Delta T \lesssim 10^{-2}$ °C). For space use, the thin-film or the Fabry-Perot filter seems preferable. The choice between the two will depend on specific requirements of narrow bandwidth and tunability.

The choice of narrowband filters for the 10.6μ wavelength is limited, since only thin-film filters with $\gtrsim 0.75$ percent fractional bandwidth presently are available. The thin-film types can be made tunable by continuous variation of the thickness of the separation layer mounted on a disc. However, from the existing infrared technology, it should be feasible to produce a narrower filter, of the Fabry-Perot type with a bandwidth on the order of a few tens of Angstroms, that will also be tunable by varying the angle of incidence.

7. OPTICAL DETECTORS

Many reports and publications have carried very elaborate discussions of the different kinds of optical detectors for various wavelengths. In this report attention will be restricted to the most promising detectors for 0.5 μ, 0.69 μ, 1.06 μ, and 10.6 μ wavelength regions, since these are the ones of greatest interest for the deep-space optical communication system considered here (Section 1).

A figure of merit frequently used for detectors[100] is $\eta^2 M^2 R$, where η and M are the quantum efficiency and internal gain, respectively, of the detector and R is the equivalent output load resistance. This figure of merit generally is bandwidth dependent and is a measure of the noise equivalent power (NEP) at the detector input. It is especially useful for comparison of wideband, or of highspeed, photodetectors when the bandwidth is fixed. The NEP is here defined* as the amount of light power which produces an rms signal current just equal to the noise current in a bandwidth of one hertz. However, the $\eta^2 M^2 R$ figure of merit is not used here to compare detectors, since the requirements for the types of mission considered in this study are general and not restricted to a specific bandwidth.

Photodetector noise also can be expressed in terms of the equivalent dc photocathode dark current $(I_D)_{Eq}$. This is determined by placing a calibrated light source at the detector input and adjusting the light input power until the rms output noise current in a 1 Hz bandwidth due to the light source is exactly equal to that due to internal detector noise. This light input power is called NEP and is related to the equivalent dc photocathode dark current $(I_D)_{Eq}$ by

$$(I_D)_{Eq} = \frac{e\eta^2}{2(h\nu)^2} (NEP)^2$$

where e is the electronic charge and $(h\nu)$ the energy per photon. It is important to distinguish $(I_D)_{Eq}$ from the dc photocathode dark current I_D (obtained by dividing the output anode dark current by the detector gain). I_D usually contains noise contributions from leakage over stem and base, dynode emission, and other sources.[101]

Comparison here is based directly on the NEP defined in units of watts/(Hz)$^{1/2}$, or detectivity (D*) defined in units of cm-(Hz)$^{1/2}$/watt. The detectivity D* is normally used to specify the noise performance of solid-state detectors and is independent of the size of the detector. NEP for a specific detector can be derived from D* using the relationship

$$NEP = \frac{\sqrt{A}}{D^*}$$

where A is the area for the detector. NEP should be such that the dominant detector noise is either background-limited (commonly called BLIP detection) or limited by the

*Note that units of this definition of NEP is watts/Hz$^{1/2}$. Also note that total noise power P_N in bandwidth B is P_N=NEP · B$^{1/2}$.

Table 33

FILTERS FOR VISIBLE AND NEAR-INFRARED WAVELENGTH REGIONS

Description	Thin-Film	Fabry-Perot	Birefringence
Bandwidth (3dB)	1.5Å at 6328Å	0.5Å at 6328Å [41] 1.6Å at 1.06μ [87]	0.125Å at 6582.8Å [94]
Blocking filter requirement	None	Yes, less than 15Å wide [91]	Yes, less than 32Å wide
Peak transmission (percent)	~30	~50 (with blocking filter)	~10 (with blocking filter)
Size: Aperture	Few inches	Few inches	~1.5 inches
Thickness	Thin (~mm)	Thin (~mm)	Few inches
Angular field of view (degrees)	~5	~2	~2
Temperature stability	~0.1Å/°C	0.05Å/°C [87]	1° retardation change per 0.01°C. Consequently, temperature stability of better than 0.01°C required.
Tunability	Yes	Yes	Yes

shot noise of the signal (noise-in-signal detection) at a detector temperature as near to 300°K as possible (that is, requiring least refrigeration).

The maximum data rate considered for the deep-space communication systems under study is on the order of one megabit, and consequently a postdetection bandwidth of 10 MHz should be adequate for the preferred direct detection at visible and 1.06-micron wavelengths (Chapter 4). Because of this relatively long response time, ~100 ns, a wide choice of available detectors exists at these wavelengths. At 10.6 microns heterodyne detection is preferred (Chapter 4) and a bandwidth of 10 MHz is adequate for heterodyne detection if correction is made for the Doppler shift prior to mixing. Without Doppler shift correction, very wideband 10.6-micron detectors (~5 GHz – see Chapter 5, Section 1.3) are needed.

In selecting detectors for comparison, an important criterion is the internal detector gain. A sufficiently large gain will decrease that portion of the equivalent noise at the input, due to thermal noise of the output circuitry, to negligible levels. Another factor in the choice of detectors is the quantum efficiency η, which should be as high as possible. It will be seen later that high gain has a stronger influence on NEP than does high quantum efficiency. For example, high-gain, low-η photomultiplier detectors provide better NEP performance at 1.06 microns than do low-gain, high-η solid-state detectors.

Some secondary considerations should be taken into account in the selection of detectors for space-flight applications. These are the photocathode area; size and weight of the detector and cooling system, if any; size and weight of the power supply for the detector; mechanical strength; and lifetime of the detectors. A comparison of detectors based on these merits depends on the individual missions, and only a general discussion is included here.

Detectors can be classified into two main categories from the point of view of the characteristics mentioned in the previous paragraph, namely (1) vacuum detectors using external photoelectric effect and (2) solid-state detectors using internal photoelectric effect. The active photosurface dimensions of the vacuum detectors can be varied from about 0.005 inch to a few inches. However, the maximum solid-state detector surface dimension is only a few millimeters and can be as low as a few tens of microns for certain wideband low-noise detectors. The small size is disadvantageous for three reasons. First, it decreases the field of view when it is used in conjunction with a collecting lens with a finite focal length. For example, a detector 50 μ in diameter, placed at the focal plane of a lens of 3 m focal length will limit the field of view to 0.015 milliradian or approximately 3 arc seconds. Second, severe mechanical alignment problems of the detector receiving telescope are also introduced. Third, it would limit the receiver diameter for an imperfect incoherent detection system.

Solid-state detectors, however, are superior to the vacuum detectors from power, weight, and size considerations. The power-supply requirements are lower for the solid-state detectors – a low-voltage power supply dissipating a fraction of a watt. However, photomultipliers need a power supply (a few kilovolts) that can dissipate a few watts and a means for thermal dissipation. The weight of a photomultiplier and its power supply will be on the order of 10 lb, an order of magnitude higher than that for a solid-state detector. Solid-state detectors also prove to be superior from considerations of ruggedness and reliability, although photomultipliers can be designed for space qualification with lifetimes sufficiently long for most missions (on the order of thousands of hours).

Sections 7.1 and 7.2 compare the NEP of various detectors at the four wavelength regions of interest. The choice between heterodyne and direct detection will be discussed in Section 7.3 in the light of the information obtained in the previous two sections. The final section contains a recommendation of the type of detectors for each wavelength.

7.1 Detectors for Visible and Near-Infrared Wavelength Regions

Because of the large current gain and low dark current, photomultipliers are attractive for the visible and near IR regions for low-level signal detection. The choice of photocathodes, with typical quantum efficiencies for the different wavelength regions, is given in Table 34.* Photomultipliers have a few hundred megahertz bandwidth capability, which is more than adequate for the present mission. Tables 35 and 36 compare detectors selected for high performance for the 0.5-μ and 0.69-μ regions, respectively. Even at room temperature, it is possible to achieve an NEP in the range of 10^{-15} to 10^{-16} watts/(Hz)$^{1/2}$. This figure can be further improved by a few orders of magnitude by cooling the detector, as shown in Tables 35 and 36. NEP data are not available for all cooled photomultipliers that have been selected for low noise. However, it is possible to estimate the NEP for these on the basis of the data available for other similar tubes. As an example, for the ITT-FW130 photomultiplier tube with S-20 photocathode listed in Tables 35 and 36, there are no published data of NEP for refrigerated conditions. An estimate for this detector can be made from the results obtained for another S-20 photocathode. It has been reported[102] that cooling an EMI-9558Q with S-20 photocathode from 300°K to 110°K will reduce the NEP by two orders of magnitude. On the basis of this experience ITT-FW130, when cooled to 110°K, can be expected to have an NEP of 1.3×10^{-18} watts/(Hz)$^{1/2}$ at 0.5 μ and an NEP of 4×10^{-18} watts/(Hz)$^{1/2}$ at 0.69 μ. Also given in Tables 35 and 36 are data for a static crossed-field photomultiplier and for the best solid-state detector, the silicon avalanche photodiode. The reason for choosing an S-20 (translucent) photocathode at 0.5-μ wavelength instead of an S-17 (opaque) photocathode with a higher quantum efficiency is that the dark current of a photomultiplier using the S-17 cathode is much higher.† From Tables 35 and 36 it can be observed that the best choice of detectors for 0.5μ and 0.69μ, from the viewpoint of NEP, are the electrostatic (no magnetic field) photomultipliers.

Table 34

CHOICE OF PHOTOEMISSIVE CATHODES FOR DIFFERENT WAVELENGTHS

Wavelength (microns)	Surface Type	Quantum Efficiency ‡ (percent)
0.53	S-17	22
0.53	S-20	20
0.6943	S-20	2.6
1.06	S-1	0.04

‡ Maximum available value for selected photocathodes. Multiple bounce photocathode structure with reflecting structures can increase the effective quantum efficiency further; see an article by Oke and Schild, Appl. Opt., April 1968, in which an increase of a factor of about 2 was reported for an S20 photocathode at 5300Å.

In the near infrared region (1.06-μ wavelength), the quantum efficiency of the photoemissive surface is reduced to 0.04 percent for even the best commercially available cathode, S-1. In contrast to this, the quantum efficiency of the photodiodes is high for the infrared spectrum and consequently makes them the apparent choice as a photodetector. However, the large current gain in photomultipliers makes their performance comparable to that of solid-state detectors, as shown in Table 37. Thus, at room temperature the photomultiplier and the two avalanche diodes all yield an NEP of about 10^{-12} watts/(Hz)$^{1/2}$. However, the photomultiplier response can be greatly improved by cooling the tube. An improvement of five orders of magnitude in NEP for Amperex 150CVP, from 10^{-11} to 10^{-16} watts/(Hz)$^{1/2}$, can be achieved[102] by cooling it from room temperature to 110°K. Using this as an estimate, one may expect to achieve an NEP of 1.2×10^{-17} watts/(Hz)$^{1/2}$ for the ITT-FW118 photomultiplier at 110°K. No NEP values this low have been reported for a solid-state detector. Thus, the photomultiplier offers substantially superior performance at 1.06-μ wavelength but at a penalty in required cooling, weight, and bulkiness.

7.2 Detection at 10.6-μ Wavelength

The choice of detector for the 10.6-μ wavelength is very limited. There are currently two principal types of detectors that meet the requirements of good sensitivity. Both require low temperature cooling, in 2°K to 77°K range, depending on detection. They are germanium extrinsic photoconductors and bolometers. The photoconductors respond faster than bolometers but have poorer sensitivity. Table 38 gives the details on three photoconductors – Ge:Cu(time constant $\tau \lesssim 1$ ns), Ge:Hg ($\tau \sim 10$ ns), and Ge:Au ($\tau \sim 30$ ns) – and two bolometers – Ge ($\tau = 400\ \mu s$)[106] and

*Table 34 is restricted to selected commercially available photocathodes; other experimental materials with higher quantum efficiency have been reported and are under development.

†A search was made for quantitive dark current data on S-17 photocathodes, but none was obtained.

Table 35

DETECTORS FOR 0.53 μ WAVELENGTH REGION

Photodetector	Quantum Efficiency (percent)	Diameter (mm)	Amplification	Temperature (°K)	N E P in Watts/ (Hz) 1/2	Reference	Comments
ITT-FW130 Photomultiplier (S-20 cathode) (FW 143 is ruggedized version)	–	2.5	2×10^6	300	1.3×10^{-16}	101	Calculation at 110°K is based on the performance on cooling other S-20 photocathodes.[101] NEP may be further improved by decreasing effective cathode size, such reduction accomplished by an image lens system.
	–	2.5	2×10^6	110	1.3×10^{-18}*		
EMI-9558 (S-20 cathode)	17	44	3.6×10^6	300	1.3×10^{-14}	102	NEP of 5×10^{-16} watts/Hz^{½} at room temperature and 5×10^{-17} watts/Hz^{½} at $\leqslant$ 235°K have been reported for selected tubes. Reference 102 reports about 20 dark counts/sec at -65°C.
	17	44	3.6×10^6	110	1.3×10^{-16}	102	
RCA-7265 (S-20 cathode)	14	40	2×10^7	300	3×10^{-14}	104	These are average of the measured NEP values of a number of selected tubes and were found to be higher than manufacturer's specifications.
	14	40	2×10^7	200	9×10^{-16}	104	
Static-Crossed Field Photomultiplier (S-20 cathode)	–	–	2×10^5	300	1.5×10^{-15}	100	The circuitry has been optimized for low NEP with 1 GHz bandwidth response.
Silicon Avalanche Photodiode	–	0.04	100	300	1.1×10^{-12}	100	The circuitry has been optimized to achieve low NEP with 1 GHz bandwidth.

*Estimated NEP and not measured.

Table 36

DETECTORS FOR 0.69μ WAVELENGTH REGION

Photodetector	Quantum Efficiency (percent)	Diameter (mm)	Amplification	Temperature (°K)	N E P in Watts/ $(Hz)^{1/2}$	Reference	Comments
ITT–FW130 Photomultiplier (S-20 cathode). (FW-143 is ruggedized version)	–	2.5	$2x10^6$	300	$4x10^{-16}$	101	Calculation at 110°K is based on the performance on cooling other S-20 photocathodes[104]. NEP may be further improved by decreasing effective cathode size, such reduction accomplished by image lens system.
	–	2.5	$2x10^6$	110	$4x10^{-18}$*		
EMI–9558 Photomultiplier (S-20 cathode)	3.8	44	$3.6x10^6$	300	$4.4x10^{-14}$	102	NEP of $1.8x10^{-15}$ watts/$Hz^{1/2}$ at room temp. and $1.8x10^{-16}$ watts/$Hz^{1/2}$ at ≤ 235°K have been reported for selected tubes. Reference 102 reports about 20 dark counts/sec at -65°C.
	3.8	44	$3.6x10^6$	110	$4.4x10^{-16}$	102	
RCA–7265 Photomultiplier (S-20 cathode)	4	40	$2x10^7$	300	$1.5x10^{-13}$	104	These are average of the measured NEP values of a number of selected tubes. The measured values were found to be higher than the manufacturer's specifications.
	4	40	$2x10^7$	200	$3.5x10^{-15}$	104	
Static-crossed field Photomultiplier (S-20 cathode)	–	–	$2x10^5$	300	$7x10^{-15}$	100	The circuitry has been optimized for low NEP with 1 GHz bandwidth response.
Silicon avalanche photodiode	–	0.04	100	300	$7x10^{-13}$	100	The circuitry has been optimized for low NEP with 1 GHz bandwidth response.

*Estimated NEP and not measured.

Table 37

DETECTORS FOR 1.06 μ WAVELENGTH

Photodetector	Quantum Efficiency (percent)	Diameter (mm.)	Amplification	Temperature (°K)	NEP in Watts/(Hz)1/2	Reference	Comments
ITT-FW118 Photomultiplier (S-1 cathode) (FW-142 is ruggedized version)	–	2.5	10^7	300	1.2×10^{-12}	101	Calculation at 110°K is based on the performance on cooling other S-1 photocathodes[101]. NEP may be further improved by decreasing effective cathode size, such reduction accomplished by image lens system.
	–	2.5	10^7	255	3×10^{-15}	101	
	–	2.5	10^7	110	1.2×10^{-17}*		
Amperex 150 CVP Photomultiplier (S-1 cathode)	0.03	32	4×10^6	300	4×10^{-11}	102	
	0.03	32	4×10^6	110	4×10^{-16}	102	
RCA 7102 Photomultiplier (S-1 cathode)	0.04	40	1.5×10^5	300	7×10^{-12}	RCA Tube Manual	
Silicon Avalanche Photodiode	–	0.04	100	300	1.5×10^{-12}	100	Circuitry optimized for low NEP with 1 GHz bandwidth response.
Germanium Avalanche Photodiode	–	0.04	25	300	1.5×10^{-12}	100	Circuitry optimized for low NEP with 1 GHz bandwidth response.

*Estimated NEP and not measured.

C (τ = 0.01 s)[107]. Specific diameters are selected to illustrate typical NEP values. The NEP for a 0.5-mm diameter Ge:Hg photoconductor is 1×10^{-12} watts/(Hz)$^{1/2}$ and for a 5-mm diameter is about 1×10^{-11} watts/(Hz)$^{1/2}$. A 4.4-mm diameter germanium bolometer has an NEP of 1.6×10^{-11} watts/(Hz)$^{1/2}$ at 4.2°K, and 5×10^{-13} watts/(Hz)$^{1/2}$ at 2.15°K.

A few words of caution are in order in interpreting Table 38. D* values of all the three photoconductors appear lower than the D* value of the background thermal radiation, which is 5×10^{10} cm-Hz$^{1/2}$/watt at 300°K for 2π steradian field of view. Thus for a direct detection system using photoconductors the background noise appears to be limiting. This is not necessarily a basic limit, since the background intensity can be reduced by several orders of magnitude by cooled narrow bandpass filters. The filter narrows down the field of view and, in addition, allows only a small portion of the black body spectrum of the background radiation to pass through, thus increasing D*. The cold shielding also reduces thermal noise from the optical components in front of the detector. Thus, even at a 10.6-μ wavelength, the noise in signal operation may be feasible using a photoconductor with cooled filter. Further development of efficient narrow band cooled filters at 10.6 μ is needed before definite conclusions can be made.

7.3 Heterodyne Vs. Direct Detection

The choice between heterodyne detection and direct detection at any of the wavelengths is dependent on many factors and is considered at length in Chapter 4, Section 9.

Discussion here will be limited to a comparison from the viewpoint of equivalent detector noise only. Background sky noise and noise in signal are not considered.

Direct detection has several practical advantages over heterodyne reception. They include:

1. Less stringent requirements on transmitter frequency stability and spectral quality
2. No local oscillator (including associated critical alignment with signal wavefront)
3. No Doppler shift correction
4. Useful detector diameter can be larger, permitting larger collector diameter and less critical detector alignment
5. Less required tracking precision
6. Useful collector diameter not limited by the finite coherence diameter dictated by atmospheric turbulence.

Tables 34 through 36, discussed previously, show the NEP in watts/(Hz)$^{1/2}$ for the different detectors based on a continuous analog input to the detector. A plot useful for application of the NEP data to a digital communication system is that of noise equivalent energy per bit N as a function of bit rate R_b. Figures 88 through 90 show these plots for 0.5, 0.69, and 1.06 micron detectors respectively, where N is computed from NEP values, bandwidth B, and bit rate R_b by the relationships:

$$N = \mathrm{NEP}\cdot B^{1/2}/R_b$$

$$B \simeq R_b$$

$$N \simeq \mathrm{NEP}/R_b^{1/2} \text{ (joules/bit)}$$

The quantum limit of 1 photon per bit also is indicated on each of the three figures. It may be noted from the figures that N decreases with increasing bit rate and extends below the quantum limit for some of the detectors at $R_b \sim 10^6$ bits/sec.

If the detector at 0.5 μ or 1.06 μ wavelength is either on a ground station or in a synchronous satellite above the Earth's atmosphere, it can be seen from Figures 88 and 90 that for an information rate of one megabit per second the photomultiplier detector noise can be below the quantum limit. If the receiving station is on the ground, direct detection will then be superior to the heterodyne detection because of degradation of beam coherence and the advantage of larger collector size that can be used. Outside the Earth's atmosphere, beam coherence degradation is negligible and the collector size will be limited by the spaceship weight and stability. Even then, it might be advantageous to use direct detection using photomultipliers. However, if the size, weight, and power advantages mentioned in Section 7.1 for a solid-state detector become dominant, heterodyne detection at 0.5 and 1.06 microns must be considered.

The 0.69-μ wavelength is attractive primarily from its high peak power capability and is considered for application as an optical beacon from the ground (see Sections 1 and 4). Figure 89 shows that, at a data rate of 10^3 or higher, it is possible to achieve an NEP for an uncooled photomultiplier within three orders of magnitude of the quantum limit. During acquisition, the angle of arrival of the beam may be far from the axial direction of the system optics, and this may cause special problems if a heterodyne system is used. Direct detection therefore appears to be most promising at this wavelength, even though the equivalent noise energy per bit is appreciably above the quantum limit.

At the 10.6 μ wavelength, the NEP for the best photoconductor at a 1-megabit rate is about 10^{15} joules/bit, which is about five orders of magnitude higher than the quantum limit of 1.87×10^{-20} joules/bit. Thus the background and detector noise is unduly excessive for direct detection at 10.6 μ, and heterodyne detection is preferred.

7.4 Conclusion

The following conclusions are made on the basis of the above discussion. Commercially available electrostatic photomultipliers have been selected on the basis of lowest

Table 38

DETECTORS FOR 10.6μ WAVELENGTH

Type	Diameter (mm)	Temperature °K	NEP Watts/(Hz)$^{1/2}$	D* cm–(Hz)$^{1/2}$/Watt	Reference	Comments
Ge:Hg.	0.25	≤ 40	0.5×10^{-12}	3.5×10^{10}	105	D* and NEP are for 60° field of view (FOV). Response time ~ 10 n sec[106]
	5.0	≤ 40	1×10^{-11}	3.5×10^{10}	105	
Ge:Cu.	0.25	≤ 14	1.7×10^{-12}	1.3×10^{10}	105	D* and NEP are for 60° FOV. Response time $\lesssim$ 1 n sec*
	5.0	≤ 14	3.5×10^{-11}	1.3×10^{10}	105	
Ge:Au.	0.25	≤ 60	0.55×10^{-9}	4×10^{7}	105	Response time ~ 30 n sec*
	5.0	≤ 60	1.1×10^{-8}	4×10^{7}	105	
	0.25	77	1.1×10^{-9}	2×10^{7}	105	
	5.0	77	2.2×10^{-8}	2×10^{7}	105	
HgTe–CdTe	~1.0	77	$\sim 10^{-11}$	$\sim 10^{10}$	106	D* and NEP are for 30° FOV. Response time < 10 n sec.
Ge Bolometer	4.4	2.15	5×10^{-13}	8×10^{11}	107	Response time = 400μ secs. D* at 4.2°K corresponds to 300°K background for 2π steradian FOV
	4.4	4.2	1.6×10^{-11}	5×10^{10}	107	
C Bolometer	5.0	2.1	1×10^{-11}	4.5×10^{10}	107	Response time = 0.01 sec

*Private communication with P.K. Cheo, Bell Telephone Laboratories.

Table 39

RECOMMENDED DETECTORS

Wavelength in μ	Type of Detector	NEP in Watts/(Hz)$^{1/2}$ or D* in cm-(Hz)$^{1/2}$ /Watt	Detector Temperature (°K)	Comments
0.5	Photomultiplier (S-20 photocathode)	10^{-16} 10^{-18}*	300 110	NEP may be improved by decreasing effective cathode size
0.69	Photomultiplier (S-20 photocathode)	4 x 10^{-16} 4 x 10^{-18}*	300 110	NEP may be improved by decreasing effective cathode size
1.06	Photomultiplier (S-1 photocathode)	10^{-12} 10^{-15} 10^{-17}*	300 255 110	NEP may be improved by decreasing effective cathode size
10.6	Photoconductor			Photoconductor performance
	(Ge: Hg)	3.5 x 10^{10} (D*)	≤40	severely limited by backgrounc
	(Ge: Cu)	1.3 x 10^{10} (D*)	4.2	Coherent detection preferred.
	Bolometer			Bolometers operate under
	(Ge)	5 x 10^{10} (D*)	4.2	BLIP condition (reached at
		8 x 10^{11} (D*)	2.15	4.2°K for this sample). However time response is slow (400 μsec).

*Estimated NEP and not measured.

noise performance, at 0.53 μ, 0.69 μ, and 1.06 μ wavelengths. The attainable NEPs for these wavelengths are summarized in Table 39. At room temperature, NEP values of the orders of 10^{-16}, 10^{-15}, and 10^{-12} watts/(Hz)^{½}can be achieved at 0.53 μ, 0.69 μ, and 1.06 μ wavelengths respectively. One or more orders of magnitude improvement in NEP may be effected by cooling the detector. An NEP of 10^{-15} watts/(Hz)^{½} has been achieved at 1.06 μ by cooling the detector to 255°K. A further decrease of two more orders of magnitude in NEP may be feasible by reducing the temperature to 110°K. It is worth noting that, even at room temperature, none of the special vacuum detectors such as the crossed-field type (with higher price and weight) or solid-state detectors (with smaller photosurface) outperform the commercially available photomultipliers in terms of low NEP for the megahertz bandwidth system. Of course, penalty has to be paid in terms of weight (~10 lb), size (~0.05 ft^3), and power dissipation (~1 watt). Background noise due to scattered sunlight can be limiting at these wavelengths; with the use of narrowband filters it can be reduced significantly at the detector surface. Thus the detection will be only shot-noise (or noise-in-signal) limited. Direct detection is preferred for 0.53 μ, 0.69 μ, and 1.06 μ wavelengths.

The threshold of detection at the 10.6-μ wavelength may be background-limited (depending on the extent of cooling and filtering used) or detector-noise limited [D* ~ 10^{11} cm-(Hz)^{½} watt for megahertz bandwidth system]. Hybrid detection can be considered at 10.6μ. The lower-noise Ge bolometer can be used for low data rate tracking channel as a direct detector [D* = 5X10^{10} cm(Hz)^{½}/watt at 4.2°K] and a cooled photoconductor as a heterodyne detector [D* – 3.5X10^{10} cm(Hz)^{½}/watt at ≤ 40° K]. These results are summaiized in Table 39.

8. OPTICAL AMPLIFIERS

Section 4.8 considers the use of an optical predetection amplifier at the receiver terminal for enhancement of the signal-to-noise ratio in direct detection systems. The analysis there shows that the amplifier gain G should be $\gtrsim 1/\eta\tau_o$ at visible and near infrared wavelengths to overcome losses due to quantum efficiency η and receiver tiansmission τ_o. Photocathode quantum efficiencies are 11 and 0.04 percent respectively at 0.5 and 1.06 microns (Section 7) which, for τ_o = 0.3, correspond to desired gain magnitudes G of 15 and 39 dB respectively. Somewhat higher gains are useful at 10.6 microns because detectors with low-noise internal gain presently are not available. A 10.6 micron predetection amplifier with ~ 50 dB gain appears advantageous to minimize effects of the high inherent noise of photoconductor detectors (Chapter 4, Section 8). It is the purpose of this section to discuss relevant optical amplifier devices; emphasis is placed on experimentally achieved gain magnitudes. Coupling of the optical radiation from the large diameter receiver mirror into the amplifier can be provided by an appropriate collimating lens near the receiver focal plane.

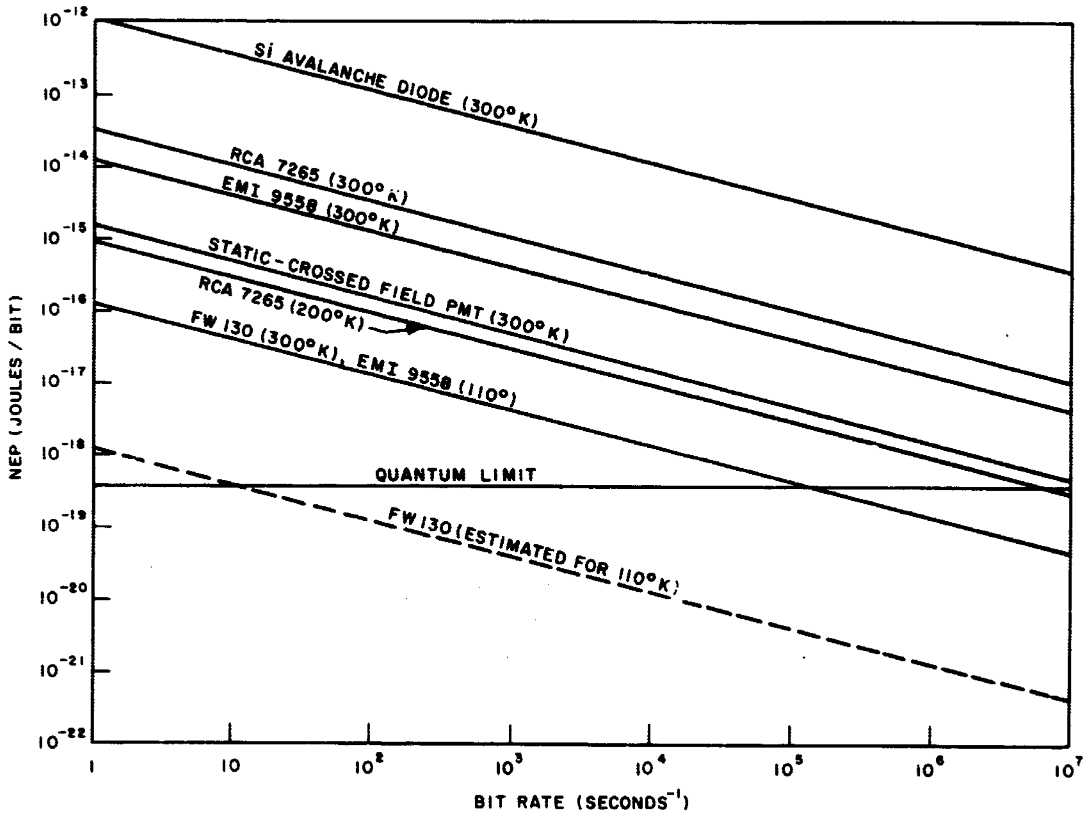


Figure 88. Noise equivalent power NEP vs. communication bit rate for detectors at 0.5 micron wavelength (see table 35)

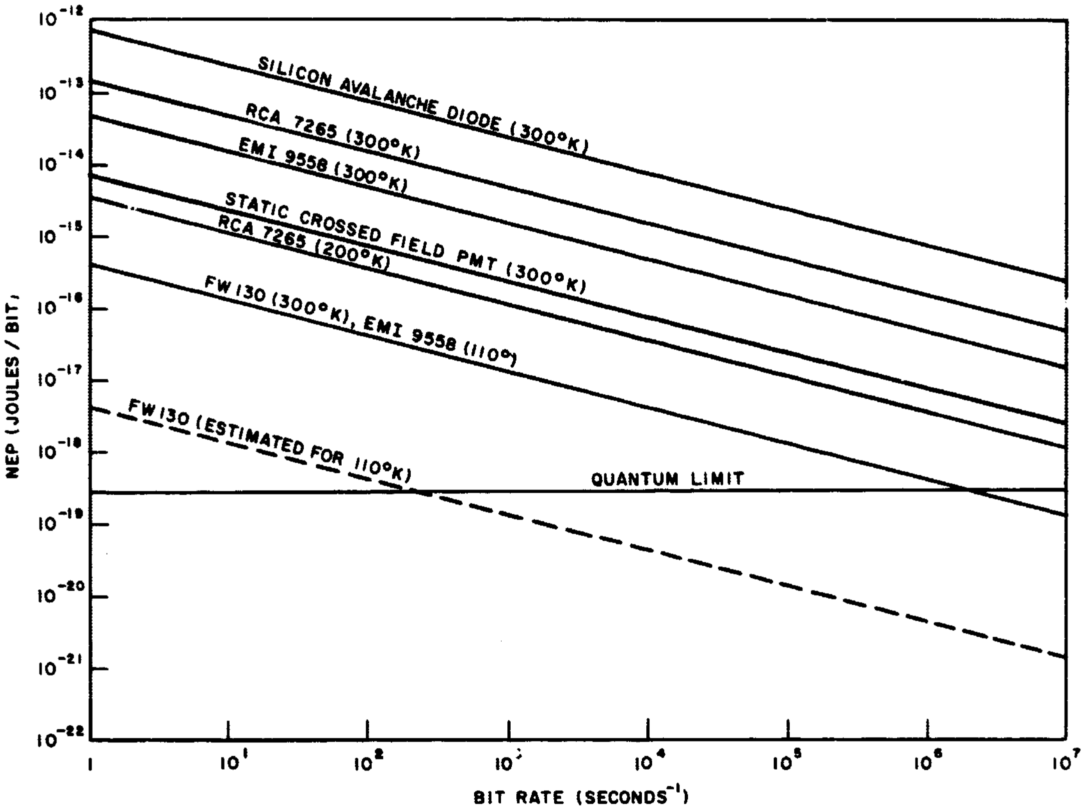


Figure 89. Noise equivalent power NEP vs. communication bit rate for detectors at 0.69 micron wavelength (see table 36)

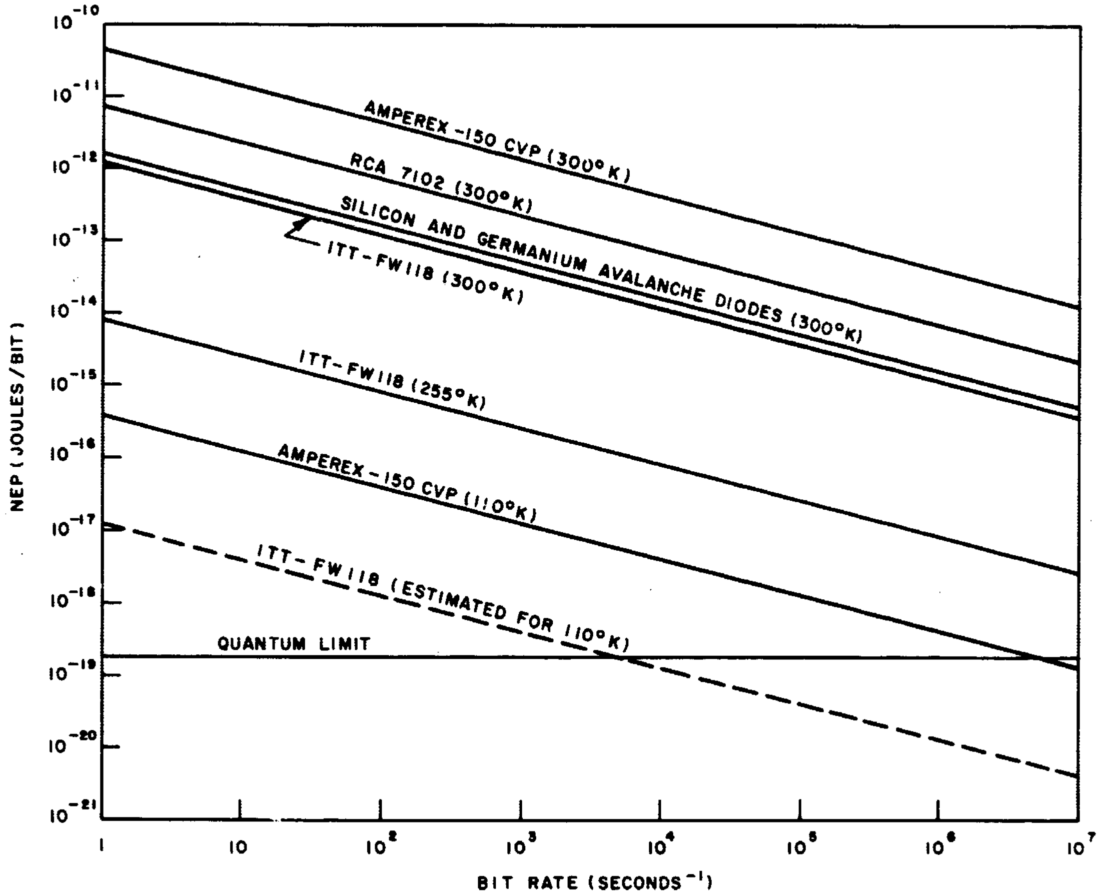


Figure 90. Noise equivalent power NEP vs. communication bit rate for detectors at 1.06 micron wavelength (see table 37)

8.1 Argon Ion Amplifier (0.48 and 0.51 Micron)

Gain measurements[110] at 4880A and 5145A of 1 and 2 mm bore argon amplifiers are shown as a function of discharge current in Figure 91. Data were obtained with no axial magnetic field; higher gain is expected with an H field, ~ factor of 2 increase, based on output power enhancement data.

Gain is much higher at 4880A, as is well known[111], even though the output powers at 4880 and 5145A are comparable. Based on the limited data of Figure 91, gain $G \propto 1/D^3$ at 4880A, which is a stronger dependence on tube bore D than the normal $G \propto 1/D^2$ observed for most neutral gas lasers. Input coupling to the amplifier must be provided with an appropriaté matching lens (see, for example, Secti. 2.1). The ratio of D to amplifier length L for optimum match is $D^2/L = \frac{4\lambda}{\pi}$ [Equation (7), Section 2]. If D = 1 mm, then the maximum possible length L at 0.5 micron is 150 cm. Typically, a factor of ten margin is recommended; i.e., $D^2/L = 40\ \lambda/\pi$, to allow for alignment tolerances and aberrations in which case L = 15 cm for D = 1 mm. The desired gain of ~ 15 dB at argon laser wavelengths appears feasible in a single pass, only for the 4880A line and for which the amplifier bore is $\lesssim$ 1 mm. The length of such an amplifier must be ~ 100 cm or more (using data of Figure 91), which requires an optical coupling geometry very close to the diffraction limit. In short, feasibility of a single-pass argon predetection amplifier is marginal. Other approaches, such as the wall stabilized arc discharge,[14] may improve the performance; e.g., a larger amplifier bore would permit multiple passes in a single discharge, but these await future developments. Power and weight requirements of the amplifier appear comparable to those of the argon laser (Section 1.3).

8.2 Nd:YAG Amplifier (0.53 and 1.06 Micron)

Gain at the 1.06 micron wavelength in Nd:YAG is quite low, ~ 7 percent in a 1-1/4 inch long by 0.2 inch diameter rod.[112] Consequently, the probability of attaining the desired 1.06 micron gain of ~ 40 dB is very unlikely with Nd:YAG, even when multiple passes in larger diameter and longer rods are considered. A possible alternative is a Nd amplifier using host crystals other than YAG. High small-signal pulse gain, 60 dB in a 1 meter long rod, has been reported in Nd:glass[113] when pumped with 24 K joules in 3 ms. CW operation of Nd at 1.06 microns at 300°K with other host crystals include[114] $CaWO_3$ (1.0580 microns) and $CaMoO_4$ (1.0610 microns). A definitive assessment on the practicality of non-YAG host crystals awaits additional data on their gain and linewidth characteristics. Nd:glass is perhaps the most promising material, but CW operation in rods with diameters sufficiently large for practical interest remains to be demonstrated. In addition, match of the 1.0648 micron Nd:YAG to the slightly shorter 1.06 + micron wavelength but broader linewidth of CW Nd:glass would have to be investigated.

Means for direct amplification of the 0.5324 micron second-harmonic wavelength of Nd:YAG is improbable because no laser transition from gaseous or solid lasers is known to exist at this wavelength. Tunable optical parametric amplifiers[115] are a possibility but appear only speculative at present for 0.5324 micron.

8.3 CO_2 Amplifier (10.6 Micron)

Gain of nonflow[116] and flowing[117-119] CO_2 amplifiers at 10.6 microns has been determined over a wide range of amplifier geometry and discharge parameters. Figure 92 shows gain G as a function of amplifier bore diameter D near optimum discharge and gas pressure conditions.[116, 119] Highest gain is obtained with a CO_2:He:N_2 mixture in the flowing case and with CO_2:He in a nonflowing (or static) amplifier. Data for the widely studied CO_2:N_2 mixture is also included to illustrate the large gain enhancement when helium is added to the flowing mixture.

A multiple-pass 10.6 micron amplifier has been described by Kogelnik and Bridges.[117] They obtained about 15 dB total small signal gain in a 20 mm diameter 5-pass amplifier. Respective discharge and cavity lengths were 1 and 1.5 meters. Higher gain (~ 20 dB in five passes) should be possible, as indicated in Figure 92, with a CO_2 flow rate of 100 cc/min STP and with a mixture of CO_2:N_2:He of 1.3:1.5.4 torr[119] instead of the 40 cc/min STP CO_2 flow rate and the higher gas pressure indicated in Reference 117. Increases in CO_2 flow rate above 100 cc/min STP can yield further gain enhancement; e.g., 7.8 dB/meter was observed in a 12 mm bore amplifier at 160 cc/min STP CO_2 flow rate.[119] A gain ~ 50 dB can be obtained by placing three such amplifiers in series with a collimating lens between each one. Gain with a non-flowing gas is much less, about 1/3 of the flowing case in a 20 mm bore amplifier. Lifetime is also less (Section 1.4). For a ground-based predetection amplifier at 10.6 microns, the flowing gas multipass type is the one to use; with a satellite receiver relay, the choice is not as clear, but gain and life advantages of the flowing type appear substantial relative to the weight advantage of the nonflowing amplifier. As stated also for the argon amplifier, weight and power requirements of each CO_2 amplifier should be very nearly the same as that of the oscillator (Section 1.3).

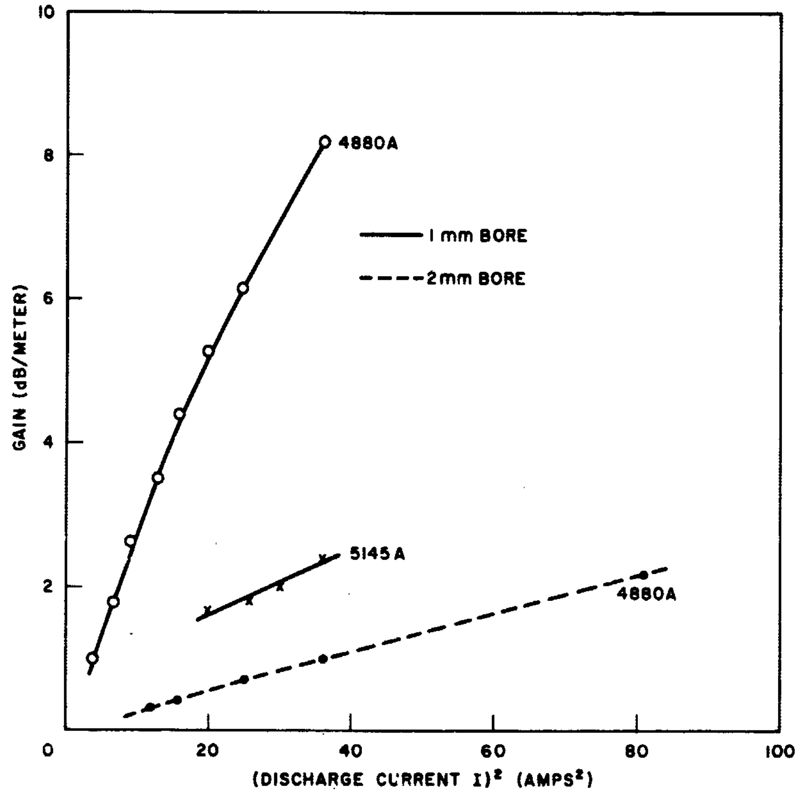


Figure 91. Single-pass low signal gain of argon ion laser amplifier vs. discharge current (reference 110)

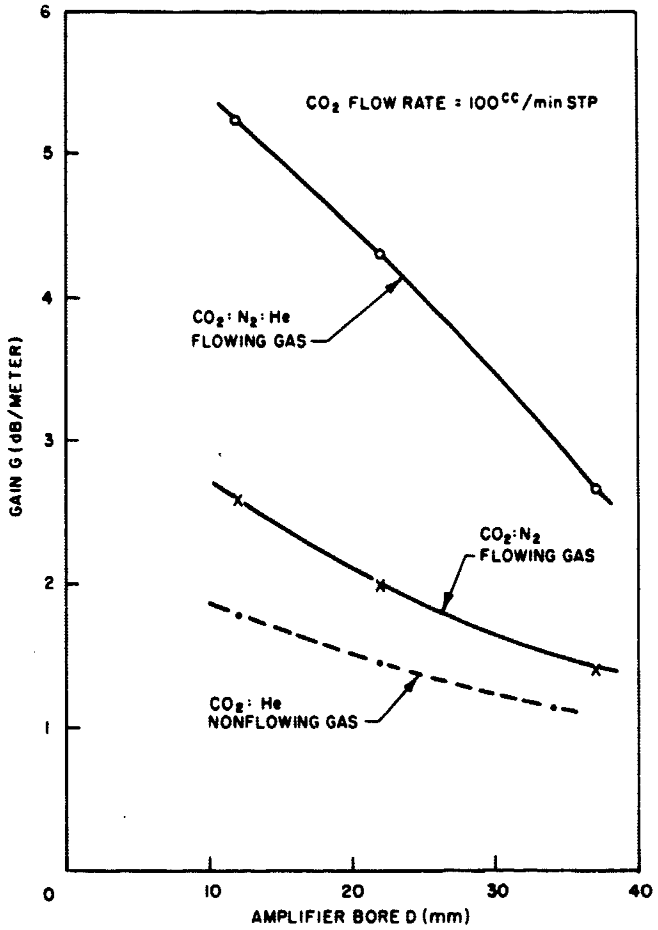


Figure 92. Single-pass low signal gain of 10.6 micron CO_2 optical amplifiers vs. amplifier bore (at optimum gas pressure and excitation conditions – references 116 and 119)

REFERENCES

1. A.L. Bloom, Appl. Opt., 5 (1966), p 1500.
2. R.G. Smith, K. Nassau, and M.F. Galvin, Appl. Phys. Letters 7 (1965), p 256.
3. J.E. Geusic, H.J. Levinstein, J.J. Rubin, S. Singh, and L.G. VanUitert, Appl. Phys. Letters, 11 (1967), p 269; private communication with J.E. Geusic.
4. Z.J. Kiss and R.J. Pressley. Appl. Opt., 5 (1966), p 1474.
5. M.I. Nathan, Appl. Opt., 5 (1966), p 1514.
6. Model 40 CO_2 Laser (with aperture to suppress higher order modes), Coherent Radiation Laboratories, Palo Alto, California.
7. E.F. Labuda, E.I. Gordon, and R.C. Miller, IEEE J. of Quantum Electronics, QE-1 (1965), p 273.
8. W.B. Bridges and A.S. Halsted. Gaseous Ion Laser Research. Final Report. Contract AF33(615)-3077, Tech. Report AFAL-TR-67-89 (May 1967).
9. R. Paananen. IEEE Spectrum, 3 (1966), p 88.
10. I. Gorog and F.W. Spong. RCA Rev. 28 (1967), p 58.
11. K.G. Hernqvist and J.R. Fendley, Jr., IEEE J. of Quantum Electronics, QE-3 (1967), p 66.
12. R.A. Knapp and R.C. Schwickert, Proceedings of Third Conference on Laser Technology. Vol. I. Office of Naval Research, Boston, Massachusetts p 131 (unclassified article within a classified volume).
13. Hughes airborne argon laser (W.B. Bridges, private communication).
14. H. Boersch et al, Physics Letters. 24A (1967), p 695.
15. A. Ashkin et al, Appl. Phys. Letters, 9 (1966), p 72; H.J. Levinstein et al, J. of Appl. Phys., 38 (1967), p 3101.
16. C. Patel, J. Chemie Phys., 64 (1967), p 82.
17. P.K. Cheo and H.G. Cooper, J. of Quantum Electronics, QE-3 (1967), p 79; also P.K. Cheo (December 1967).
18. T.F. Deutsch, IEEE J. of Quantum Electronics, QE-3 (1967), p 151.
19. T.J. Bridges and C.K.N. Patel, Appl. Phys. Letters, 7 (1965), p 244.
20. W.B. Bridges, P.O. Clark, and A.S. Halsted, High Power Gas Laser Research, Final Report, Contract AFAL-TR-66-369 (January, 1967).
21. J.E. Geusic, H.M. Marcos, and L.G. VanUitert, Proceedings of the 1965 Physics of Quantum Electronics Conference, P. Kelley, B. Lax, and P. Tannenwald, Editors, (New York, McGraw-Hill, 1966); J.E. Geusic (to be published).
22. Private communication with A. Bloom.
23. Private communication with D. MacNair.
24. Private communication with W. Witteman.
25. Private communication with H. Nu, Systems, Inc. Palo Alto, California.
26. Private communication with I.T. Basil, Westinghouse-Baltimore, Surface Division.
27. S.E. Harris, Appl. Optics, 5, (1966) p 1639.
28. Private communication with R.G. Smith.
29. I.P. Kaminow and E.H. Turner, Appl. Opt., 5 (1966), p 1612.
30. A. Ashkin et al, Appl. Phys. Letters, 9 (1966), p 72; H.J. Levinstein et al, J. of Appl. Phys., 38 (1967), p 3101.
31. J.E. Geusic, H.J. Levinstein, J.J. Rubin, S. Singh, and L.G. VanUitert, Appl. Phys. Letters, 11 (1967), p 269.
32. R.T. Denton, F.S. Chen, and A.A. Ballman, J. of Appl. Phys., 38 (1967), p 1611.
33. T.E. Walsh, RCA Review, 27 (1966), p 323.
34. A. Yariv, C.A. Mead, and J.V. Parker, IEEE J. of Quant. Elect., QE-2 (1966), p 243.
35. M. Teich and T. Kaplan, IEEE J. of Quant. Elect., QE-2 (1966), p 702.
36. C.F. Quate, C.D.W. Wilkinson, and D.K. Winslow, Proc. IEEE, 53 (1965), p 1604.
37. E.I. Gordon, Appl. Opt., 5 (1966), p 1629.
38. R. Adler, IEEE Spectrum, 4 (1967), p 42.
39. R.W. Dixon, J. of Appl. Phys. (to be published).
40. N.F. Foster and G.A. Rozgonyi, Appl. Phys. Lett., 8 (1966), p 221.
41. R.W. Dixon, IEEE J. of Quant. Elect., QE-3 (1967), p 85.
42. R.W. Dixon and A.N. Chester, Appl. Phys. Lett., 9 (1966), p 190.
43. Ground-Based Astronomy – A Ten Year Program (Washington, D.C., National Academy of Sciences – National Research Council, 1964).
44. Crawford, D.L. Editor, The Construction of Large Telescopes (New York, Academic Press, 1966).
45. M. Glos, Scientific Research (June 1967), p 45.
46. The New York Times (June 11, 1967).
47. I.S. Bowen, The Astronomical Journal, 69 (1964), p 816.
48. R.N. Watts, Sky & Telescope, 28 (1964), p 78.
49. Private communication with H. Wischnia, Perkin-Elmer Corporation.
50. W.A. Cosby and R.G. Lyle, The Meteoroid Environment and Its Effects on Materials and Equipment, NASA SP-78 (Washington D.C., U.S. Government Printing Office, 1965).

51. C.T. D'Aiutolo, Nucleonics, 21 (1963), p 51.

52. W.M. Alexander et al, Science, 149 (1965), p 1240.

53. R.A. Becker, Appl. Opt., 6 (1967), p 955.

54. F.L. Whipple and E.L. Fireman, Nature, 183, (1959) p 1315.

55. L.D. Jaffe and J.B. Rittenhouse, ARS Journal, 32, (1962) p 320.

56. L. Spitzer, Jr., The Astronomical Journal, 65, (1960) p 242.

57. R.M. Scott, Appl. Opt., 1, (1962) p 387.

58. M.S. Lipsett et al, Laser/Optics Techniques, 2nd Interim Summary Report, Perkin-Elmer Report No. 8631 (December 31, 1966). This is the latest in a series of such reports.

59. R.E. Danielson, International Science and Technology (July 1967), p 54.

60. L. Spitzer, Jr. and Bruno A. Boley, J. Opt. Soc. Am., 57 (1967), p 901.

61. G.P. Kuiper and B.M. Middlehurst, editors, Stars and Stellar Systems, Vol. 1, "Telescopes" (Chicago, The University of Chicago Press, 1960).

62. R.W. Dietz and J.M. Bennett, Applied Optics, 6 (1967), p 1275.

63. Private communication with H. Tschunko.

64. Private communication with M. Schwarzschild, Princeton University.

65. J.M. Burch, Nature, 171 (1953), p 889.

66. E.H. Eberhardt, ITTIL Applications Note E3 (December 6, 1963).

67. A.E. Lopez et al, J. Spacecraft, 1 (1964), p 399.

68. Private communication with G. Allen Smith, NASA, Ames, Iowa.

69. J. Buckley et al, Giant Aperture Telescope Study Phase II Report, Perkin-Elmer Report No. 8558 (November 11, 1966).

70. Private communication with A.B. Meinel, U. of Arizona.

71. Private communication with J.S. Courtney-Pratt.

72. Private communication with R.A. Becker, Sacramento Peak.

73. E.R. Schlesinger, Electronics (February 8, 1963), p 47.

74. M. Subramanian and J.A. Collinson, BSTJ, Vol. 44 (1965), p 543.

75. Private communication with M. Greenstein, Mount Palomar Observatory.

76. J.D. Buckley et al, Giant Aperture Telescope Study, Phase III Report, Perkin-Elmer Report No. 8691 (February 20, 1967).

77. Optical Space Communications System Study, Final Report, Contract NAS W-590, General Electric Co. (March 3, 1964).

78. H.E. Rowe, Lenses with Random Imperfections, to be published.

79. S.O. Rice, "Reflection of Electromagnetic Waves from Slightly Rough Surfaces," Commun. on Pure and Appl. Math, 4 (August, 1951), pp 351-378.

80. J. Ruze, "Antenna Tolerance Theory – A Review," Proc. IEEE, Vol. 54 (April, 1966), pp 633-640.

81. L.A. Chernov, Wave Propagation in a Random Medium, R.A. Silverman, Translator (New York, McGraw-Hill Book Co., 1960).

82. Private communication with F. Penniman, Quartermaster, R&D Command, Natick, Massachusetts.

83. F. Daniels, Direct Use of the Sun's Energy.

84. T. Sakurai et al, Solar Energy, 8 (1964), p 117.

85. Private communication with R.F. Lucy et al.

86. R.W. Jones et al, Giant Aperture Telescope Study, Phase I Report, Perkin-Elmer Report No. 8393 (May 12, 1966).

87. Samuel J. Holmes,"Design and Fabrication of Optical Filters for Selected Laser Frequencies," Technical Report AFAL TR-66-400 (December 1966), prepared by Spectrolab for Wright-Patterson Air Force Base, Dayton, Ohio.

88. B. Lyot, Ann. Astrophys (1944), p 31.

89. S.E. Harris, E.O. Amman, and I.C. Chang, "Optical Network Synthesis Using Birefringent Crystals I. Synthesis of Lossless Networks of Equal-Length Crystals," JOSA, 54 (October 1964), pp 1267-1279.

90. E.O. Amman and Y.M. Yarborough, Birefringent Devices, Report No. NAS8-20570, prepared by Sylvania Electronic Systems, Mountain View, California, for NASA, Hunstville, Alabama.

91. J.W. Evans, "A Birefringent Monochromator for Isolating High Order in Grating Spectra," Appl. Opt., 2 (February 1963), pp 193-197.

92. E.O. Amman, "Modification of Devices Normally Operating between Input and Output Polarizers to Allow Their Use with Arbitrarily Polarized Light," JOSA, 55 (April 1965), p 413.

93. P.H. Lissberger, "Properties of All-Dielectric Interference Filters I and II," JOSA, 49 (1959), p 121.

94. M.L. Baker and V.L. Yen, "Effects of the Variation of Angle of Incidence and Temperature on Infrared Filter Characteristics," Appl. Opt., 6 (August 1967), p 1343.

95. Private communication with D.R. Herriott.

96. W.H. Steel, R.N. Smartt, and R.G. Givanelli, "A 1/8Å Birefringent Filter for Solar Research," Australian J. Phys., 14 (1961), pp 201-211.

97. Private communication with E.O. Amman, Sylvania Electronic Systems, Mountain View, California.

98. Private communication with G. Francisco, Optical Coating Laboratories, Inc., Santa Rosa, California.

99. Manufactured by Optical Coating Laboratories, Inc., and termed "Circular Variable Filter."

100. L.K. Anderson and B.J. McMurtry, "High Speed Photodetectors," Appl. Opt., 5 (October 1966), pp 1573-1587.

101. E.H. Eberhardt, "Multiplier Phototubes for Single Electron Counting," Elec. Comm., 40, No. 1 (1965) pp 214-133.

102. D.H. Olson and D.O. Landen,"A Cryostat for Optimizing Signal Noise of Photomultipliers," private communication.

103. J.P. Rodman and H.J. Smith, "Tests of Photomultipliers for Astronomical Pulse Counting Applications," Appl. Opt., 2 (February 1963), pp 181-186.

104. H.F. Wischnia, H.S. Hemstreet, and J.G. Atwood, "Determination of Optical Technology Experiments for a Satellite," Perkin-Elmer Report, NASA-CR252, prepared for NASA under Contract No. NAS8-11408.

105. Santa Barbara Research Center Catalog No. 67 CM.

106. Societé Anonyme de Telecommunications, Paris, France, Catalog DC2070B.

107. F.J. Low, "Low-Temperature Germanium Bolometer," JOSA, 51 (November 1961), pp 1300-1304.

108. W.S. Boyle and K.F. Rodgers, "Performance Characteristics of a New Low-Temperature Bolometer," JOSA, 49 (January 1959), pp 66-69.

109. Single Photoelectron Counting with Multiplier Phototubes, Application Note E5, ITT Industrial Laboratories.

110. Private communication with R.C. Miller.

111. A.L. Bloom, Appl. Opt., 5 (1966), p 1500.

112. Private communication with R.G. Smith.

113. E. Snitzer, Appl. Opt., 5 (1966), p 1487.

114. Z.J. Kiss and R.J. Pressley, Appl. Opt., 5 (1966), p 1474.

115. R.W. Minck, R.W. Terhune, and C.C. Wang, Appl. Opt., 5 (1966), p 1595.

116. P.K. Cheo and H.G. Cooper, IEEE J. of Quant. Elect., QE-3 (1967), p 79.

117. H. Kogelnik and T.J. Bridges, IEEE J. of Quant. Elect., QE-3 (1967), p 95.

118. T.F. Deutsch, IEEE J. of Quant. Elect., QE-3 (1967), p 151.

119. P.K. Cheo, IEEE J. of Quant. Elect. (December 1967).

PRECEDING PAGE BLANK NOT FILMED.

CHAPTER 4. OPTICAL COMMUNICATIONS

1. BACKGROUND NOISE

Methods of communicating optically are, to varying degrees, vulnerable to and degraded by the presence of the background radiation superposed on the signal. This vulnerability is especially important when detecting weak signals by the direct detection (i.e., energy detection) method. As is well known, background radiation is usually a much smaller handicap in heterodyne detection. This section identifies the most important sources of background radiation, presents numerical values of the background that may be considered typical, and discusses briefly the relevant statistics of the radiation. The discussion is confined to gross characterization of the background, and the many effects which determine the details of the background are neglected.

1.1 Sources and Magnitude of Background

In the visible portion of the optical region, at wavelengths below 1 micron, scattered sunlight is the most significant source of background radiation. In daytime, for a telescope on the ground (and except in the practically prohibited situation in which the telescope views a portion of the Sun), the background is due to scattering from the Earth's atmosphere. This is a diffuse background. The Sun itself may be approximated as a blackbody radiator at 6000°K, and the scattered light which reaches the telescope has a spectral distribution determined by this temperature (but smaller in magnitude than direct sunlight). For the night sky (and in the visible region) the principal background radiation is caused by sunlight scattered from extra-terrestrial objects (e.g., planets) which are unavoidably within the telescope field of view during a communication mission. The same will be true for a receiver placed outside the Earth's atmosphere on a satellite. This can be illustrated by background light from scattering of sunlight from Mars. In such a case the background may not come from the entire field of view of the telescope, since the scattering object will in general not subtend a larger angle than the field of view of the telescope. This depends on the distance to the scattering object (as well as its size) and the quality of receiver optics.

At longer wavelengths, in the vicinity of 10 microns, the radiation from direct blackbody emission will be considerably greater than that from scattered sunlight. Hence, for a receiver on the ground, the background (again diffuse) is due to blackbody emission at about 300°K. For a satellite receiver the background (at infrared) is mainly due to direct emission from the extraterrestrial body; e.g., for Mars a 238-degree blackbody.[1]

In the diffuse case the background may be usefully described by the spectral irradiance, N_λ, defined as the power per unit area-solid angle-wavelength (for example: watts/cm -steradian-micron). Alternatively, one may define the power per unit area-solid angle-Hertz, N_ν. These are simply related by

$$N_\nu = N_\lambda \frac{\lambda}{\nu} \qquad (1)$$

where λ is the wavelength and ν the frequency. Backgrounds that are not diffuse may be described by the power per unit area per wavelength (or per Hertz) at the receiver. That is, one can define the intensity spectrum π_ν or the equivalent π_λ, where

$$\pi_\nu = \pi_\lambda \frac{\lambda}{\nu} \qquad (2)$$

Another description for the same information encompasses both the diffuse and nondiffuse cases and is somewhat more directly useful for consideration of heterodyne detection as well as certain aspects of direct detection. In this description the background is assigned to a discrete number of distinguishable transverse field modes[2] (or directions of propagation) and longitudinal modes, as well as to one of the two independent polarizations with which a plane electromagnetic wave can be specified. Then the background radiation can be described either in terms of energy per mode or by the number of photons per mode, known as the degeneracy* and here denoted by δ.

For a diffuse background, where the radiation in a bandwidth W is spatially and uniformly spread out and falls on a receiver of collector area A_R with solid angle of acceptance Ω_R, the number of modes received in time t is

$$\frac{A_R \Omega_R}{\lambda^2} \quad Wt$$

hence the number of photons per mode is

$$\delta = \frac{1}{2}\left(\frac{1}{h\nu}\right)\frac{N_\nu A_R \Omega_R Wt}{\frac{A_R \Omega_R}{\lambda^2} \quad Wt}$$

or

$$N_\nu = 2\frac{h\nu}{\lambda^2} \cdot \delta \qquad (3)$$

The factor 2 arises from the convention adopted here that N_ν is for arbitrary polarization (which includes two independent polarizations), while the term degeneracy will be reserved here for only one.

For an object which does not fill the receiver field of view, there is a related expression for π_ν. Supposing that the luminous object, whether it be a scatterer or a primary emitter, subtends solid angle Ω; there are then $\Omega(\Delta A)/\lambda^2$ transverse modes in which radiation falling on the area ΔA is divided (this light is again emitted in all directions from the object), leading to the intensity spectrum:

$$\pi_\nu = 2\frac{h\nu}{\lambda^2} \cdot \delta \cdot \Omega \qquad (4)$$

For a blackbody at temperature T the degeneracy is given by

$$\delta_{\text{blackbody}} = (e^{h\nu/kT} - 1)^{-1} \qquad (5)$$

This expression is directly applicable to direct emission, and is used here for infrared wavelengths. In the visible region, one must use the effective degeneracy instead, because of the scattering of the sunlight before it reaches the telescope. For sunlight scattered by the atmosphere, the blackbody degeneracy is multiplied by a number which is the product of a scattering coefficient (of the order of 10^{-1}) and a geometric factor which is nominally the ratio of the solid angle subtended by the primary source (the Sun) to that into which the scattering takes place, of the order of 10^{-5}. Based partly on experimental observations and partly on calculation†[1]

atmosphere-scattered sunlight

$$\delta \approx 2\text{x}10^{-6}\ (e^{h\nu/6000k} - 1)^{-1}$$

If the sunlight is scattered by Mars, the effective degeneracy results from the same sort of modification to the degeneracy of the Sun, except that the scattering coefficient may differ and the solid angle of interest (that of the primary source at the scatterer) is that which the Sun subtends at Mars. A representative figure is[4]

sunlight scattered by Mars

$$\delta \approx 1.3\text{x}10^{-6}\ (e^{h\nu/6000k} - 1)^{-1}$$

In Table 40 the effective degeneracies in the background radiation are summarized for some wavelengths of interest.

By using Table 40, the amount of background radiation which falls on a receiver may be simply calculated from Equations (1) through (4). Care must be exercised to use the appropriate solid angle to calculate the contribution from the "small" source (here illustrated by Mars) which may unavoidably be within the receiver field of view in accordance with the communication mission under study.

1.2 Background Noise Statistics

In a classical or wave-like description of the electromagnetic field associated with the background noise, the electric and magnetic fields associated with a single mode of the radiation each have a Gaussian distribution, independent of one another and of the fields from every other mode of the radiation. This is the most random (or maximum entropy) distribution, which is appropriate to thermal noise. Since the energy is quadratic in the field quantities, the energy of the i^{th} mode may be written as

$$E_i = x_i^2 + y_i^2 \qquad (6)$$

where x_i and y_i are random variables which stand for electric and magnetic fields, each Gaussian, while the energy associated with N modes is

$$E = \sum_{i=1}^{N} (x_i^2 + y_i^2) \qquad (7)$$

Energy is distributed among the different modes in such a way as to yield the largest amount of randomness.

*There is a 1:1 correspondence between a mode and a cell of phase space having volume h^3 where h is Planck's constant.[3]

†This calculation ignores the fact that the scattering is frequency dependent. When applied over a narrow range of wavelengths, say from 0.5 to 1 micron, this does not result in a gross error.

section. This review is not intended to be exhaustive, but it is hoped that it presents representative data and accurately reflects the current level of understanding of these phenomena. Following this summary, an assessment of the importance of the known results to various types of optical communications systems will be presented. Some experiments are suggested which could provide much needed data to further aid in the evaluation of proposed systems.

2.1 Summary of Observed Phenomena

Some observational data describing the atmospheric influence on field amplitude, phase, and mutual coherence functions are summarized below. Depolarization will not be considered, since atmospheric depolarization at optical wavelengths has been shown to be negligible.[8]

2.1.1 Amplitude Related Effects

2.1.1.1 Attenuation. Observations of attenuation in clear air have been reported, as well as attenuation resulting from precipitation of various kinds.

A series of measurements were performed recently in Colorado by A. L. Buck,[9] and clear air attenuations along elevated propagation paths averaging approximately 0.2 dB/km have been reported. Clear air attenuation is expected to decrease rapidly with increasing altitude, however.[11]

T. S. Chu[10] has measured the attenuation due to precipitation (rain, fog, or snow) over a 2.6 km path at Bell Telephone Laboratories' Crawford Hill location in New Jersey. Data were taken at 0.63, 3.5, and 10.6μ. Attenuations from dense fog were observed in excess of 65, 45, and 60 dB above clear weather levels for the respective wavelengths. True attenuations are unknown in these instances since the latter bounds were set by receiver sensitivities. Results of measurements made in a light fog are shown in Figure 93. The attenuation is greatest at 0.63μ and is due primarily to scattering. Progressing to longer wavelengths, the attenuation drops off, and at 10.6μ absorption is the dominant effect.

Attenuation from rain does not vary as greatly over the range of wavelengths examined. Attenuation is smallest at 0.63μ and is due again primarily to scattering. Rain attenuation increases with wavelength (Figure 93), and at near infrared (IR) wavelengths, scattering and absorption contribute about equally.

No consistent wavelength dependence was observed in attenuation by snow. For the same equivalent liquid water density, attenuation from snow lies somewhere between the corresponding values for rain and fog.

2.1.1.2 Fluctuations. The term "fluctuation" generally refers to a fluctuation in the field that is deduced from an observed fluctuation in the output of an optical receiver. Two types of receivers have been used to observe optical propagation effects: direct detection and heterodyne receivers. One should exercise care in distinguishing between the two types of observations, since these devices respond to fundamentally different (but assuredly related) moments of the aperture field.

Temporal records of the fluctuating outputs of these two types of receivers appear to be very similar. They are characterized by a relatively steady noise-like trace interspersed with sharp spikes or bursts; these pulses of peak output may occur at 50 to 100 ms intervals and may be of 10 to 25 ms duration. In both cases, these peaks have been observed to rise to a limiting value under relatively good atmospheric conditions, but seem to be randomly distributed under conditions of high turbulence. The power spectra of direct detected and heterodyned signals have been observed to display differences, although the gross behavior under varying atmospheric conditions is again similar.

2.1.2 Fluctuation Magnitude

Direct detection measurements of these fluctuations and certain parametric dependences have been made by Peteranecz and Simons,[12] P. B. Taylor,[13] T. S. Chu,[14] R. F. Lucy et al,[15] M. Subramanian and J. A. Collinson,[16] and D. L. Fried et al.[17] All measurements of fluctuations in the outputs of direct detection receivers described below were made at 0.63μ.

Peteranecz and Simons[12] and P. B. Taylor[13] have described measurements over a 16.5 km path in Dayton, Ohio. The path was elevated (about 70 m on the average) and passed over many types of terrain. A 4.3 cm collecting aperture was employed and peak receiver outputs were observed to range up to 30 dB above the quasi-steady level. Simultaneous microwave measurements over the same path showed power scintillations of the order of 1 dB.

T. S. Chu[14] has observed fluctuations induced by clear air turbulence over the same 2.6 km path mentioned earlier.[10] He reports rms fluctuations of the signal averaging 30 percent of the average level, but ranging between extremes of 10 and 90 percent. The largest fluctuations were observed in the presence of the Sun. A measurement in a strong crosswind (transverse to the propagation path, with velocity ~30 mph) showed no increase in the fluctuation. An eight-fold increase in receiver aperture diameter resulted in a reduction in percent rms fluctuation by a factor of nearly two (from 60 to ~ 30 percent). Diurnal variations in fluctuation were also noted.

Buck[9] has also investigated the fluctuation observed with a direct detection receiver and he suggests the fine structure of the fluctuation records is a result of "internal breakup" of the beam, while the large bursts are due to the beam "wandering" in and out of the collector aperture. He section. This review is not intended to be exhaustive, but it is hoped that it presents representative data and accurately reflects the current level of understanding of these phenomena. Following this summary, an assessment of the importance of the known results to various types of optical communications systems will be presented. Some experiments are suggested which could provide much needed data to further aid in the evaluation of proposed systems.

2.1 Summary of Observed Phenomena

Some observational data describing the atmospheric influence on field amplitude, phase, and mutual coherence functions are summarized below. Depolarization will not be considered, since atmospheric depolarization at optical wavelengths has been shown to be negligible.[8]

2.1.1 Amplitude Related Effects

2.1.1.1 Attenuation. Observations of attenuation in clear air have been reported, as well as attenuation resulting from precipitation of various kinds.

A series of measurements were performed recently in Colorado by A. L. Buck,[9] and clear air attenuations along elevated propagation paths averaging approximately 0.2 dB/km have been reported. Clear air attenuation is expected to decrease rapidly with increasing altitude, however.[11]

T. S. Chu[10] has measured the attenuation due to precipitation (rain, fog, or snow) over a 2.6 km path at Bell Telephone Laboratories' Crawford Hill location in New Jersey. Data were taken at 0.63, 3.5, and 10.6μ. Attenuations from dense fog were observed in excess of 65, 45, and 60 dB above clear weather levels for the respective wavelengths. True attenuations are unknown in these instances since the latter bounds were set by receiver sensitivities. Results of measurements made in a light fog are shown in Figure 93. The attenuation is greatest at 0.63μ and is due primarily to scattering. Progressing to longer wavelengths, the attenuation drops off, and at 10.6μ absorption is the dominant effect.

Attenuation from rain does not vary as greatly over the range of wavelengths examined. Attenuation is smallest at 0.63μ and is due again primarily to scattering. Rain attenuation increases with wavelength (Figure 93), and at near infrared (IR) wavelengths, scattering and absorption contribute about equally.

No consistent wavelength dependence was observed in attenuation by snow. For the same equivalent liquid water density, attenuation from snow lies somewhere between the corresponding values for rain and fog.

2.1.1.2 Fluctuations. The term "fluctuation" generally refers to a fluctuation in the field that is deduced from an observed fluctuation in the output of an optical receiver. Two types of receivers have been used to observe optical propagation effects: direct detection and heterodyne receivers. One should exercise care in distinguishing between the two types of observations, since these devices respond to fundamentally different (but assuredly related) moments of the aperture field.

Temporal records of the fluctuating outputs of these two types of receivers appear to be very similar. They are characterized by a relatively steady noise-like trace interspersed with sharp spikes or bursts; these pulses of peak output may occur at 50 to 100 ms intervals and may be of 10 to 25 ms duration. In both cases, these peaks have been observed to rise to a limiting value under relatively good atmospheric conditions, but seem to be randomly distributed under conditions of high turbulence. The power spectra of direct detected and heterodyned signals have been observed to display differences, although the gross behavior under varying atmospheric conditions is again similar.

2.1.2 Fluctuation Magnitude

Direct detection measurements of these fluctuations and certain parametric dependences have been made by Peteranecz and Simons,[12] P. B. Taylor,[13] T. S. Chu,[14] R. F. Lucy et al,[15] M. Subramanian and J. A. Collinson,[16] and D. L. Fried et al.[17] All measurements of fluctuations in the outputs of direct detection receivers described below were made at 0.63μ.

Peteranecz and Simons[12] and P. B. Taylor[13] have described measurements over a 16.5 km path in Dayton, Ohio. The path was elevated (about 70 m on the average) and passed over many types of terrain. A 4.3 cm collecting aperture was employed and peak receiver outputs were observed to range up to 30 dB above the quasi-steady level. Simultaneous microwave measurements over the same path showed power scintillations of the order of 1 dB.

T. S. Chu[14] has observed fluctuations induced by clear air turbulence over the same 2.6 km path mentioned earlier.[10] He reports rms fluctuations of the signal averaging 30 percent of the average level, but ranging between extremes of 10 and 90 percent. The largest fluctuations were observed in the presence of the Sun. A measurement in a strong crosswind (transverse to the propagation path, with velocity ~30 mph) showed no increase in the fluctuation. An eight-fold increase in receiver aperture diameter resulted in a reduction in percent rms fluctuation by a factor of nearly two (from 60 to ~ 30 percent). Diurnal variations in fluctuation were also noted.

Buck[9] has also investigated the fluctuation observed with a direct detection receiver and he suggests the fine structure of the fluctuation records is a result of "internal breakup" of the beam, while the large bursts are due to the beam "wandering" in and out of the collector aperture. He

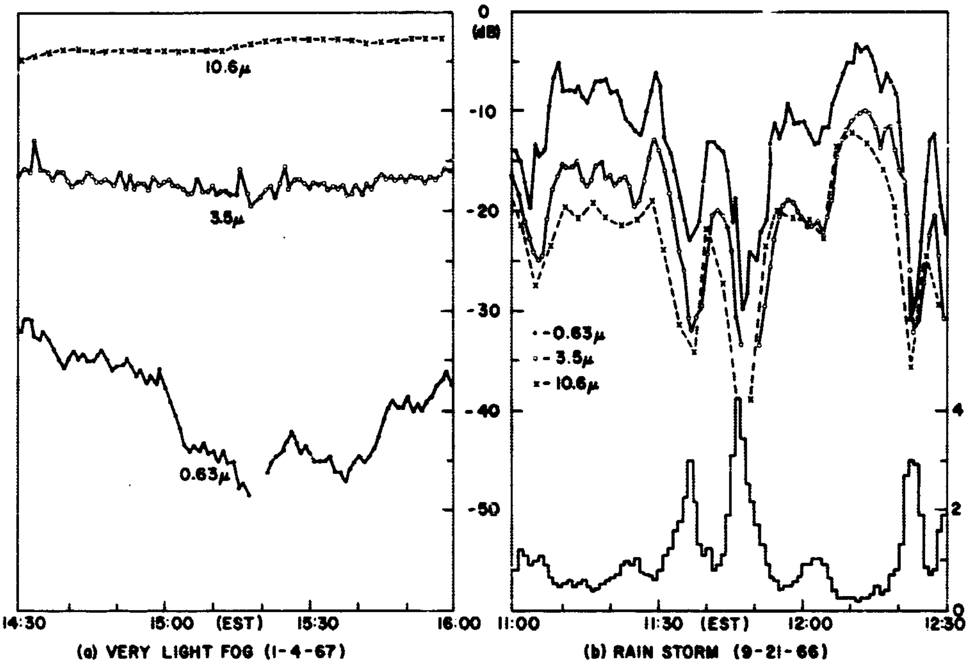


Figure 93. Measurement of 2.6 km transmission loss relative to signal level in clear weather (0 dB)

has measured the fluctuation index σ_n (defined as the ratio of sample standard deviation to sample mean) of the receiver output for a path length of 4 km as a function of receiver aperture diameter. Like Chu, he observes a definite reduction in σ_n as the receiving aperture is enlarged. Buck observed no systematic variation in σ_n for path lengths ranging from 550 m to 145 km.

Lucy and his coworkers[15] have performed simultaneous heterodyne and direct detection measurements over a 1 km path at the Sylvania Research Laboratory in Waltham, Massachusetts. In computing sample probability density functions for the direct detected signal, fluctuations were observed to be greatest and the average level lowest i.1 clear weather. The fluctuation decreased and the average level increased under weather conditions which progressed from clear to overcast to light rain. Reported observations[15] taken with different sized collection apertures appear to indicate that the percent fluctuation (or fluctuation index) decreases with increasing aperture diameter. However, Lucy indicates[18] that, considering additional (unpublished) data, no definite conclusion can be reached from his group's results.

Subramanian and Collinson,[16] at Bell Laboratories, have investigated the percent fluctuation (or "depth of modulation" as they describe it) over ranges of about 30 to 800 m. They observe a rapid decrease in fluctuation with increasing aperture size. However, they report that the percent fluctuation does not decrease beyond a finite "residual" level reached as the collecting aperture becomes larger than the direct beam. Residual fluctuations of a fraction of a percent are observed at the distances examined. When all of the beam is collected, the percent fluctuation was observed to increase with the 3/2 power of propagation distance (from 30 to 800 m). It was observed that this dependence on path length was independent of weather conditions; however, for a fixed distance the percent fluctuation depends sensitively on atmospheric conditions, having been observed to change by a factor of two or three in a matter of seconds. Attempts to determine the dependence of percent modulation on atmospheric variables were unsuccessful.

Fried, Mevers, and Keister[17] measured fluctuations at 0.63μ over an 8 km path at the Autonetics Electro-Optical Laboratory in California. Their purpose was to measure the fluctuation statistics as the size of the collection aperture was varied. The theoretically predicted log normal distribution was expected for small apertures and a transition to Gaussian statistics was expected as the receiver cross section increased. However, the statistics were observed to be log normal for all apertures varying in diameter from 1 mm to 1 m. This apparently anomalous result gave rise to the conjecture by these investigators that this phenomenon is associated with the finite cross section of the laser beam. D.H. Höhn[21,22] has also measured the statistics of intensity fluctuation at 0.63μ over 4.5 and 14.5 km paths at the University of Tübingen, Germany. He observed the log normal distribution for small collection apertures, but has also observed deviations therefrom. In particular, he has recorded normal distributions for the intensity for the largest (80 mm) apertures used.

Measurements of the fluctuation in the output of a heterodyne receiver have been made by Rosner[19] and also by Lucy et al.[15] These measurements were performed at 0.63μ. F.E. Goodwin[20] has also recently reported observing fluctuations in a 3.39μ heterodyne system.

Rosner utilized the same 2.6 km propagation path employed by Chu at Bell Laboratories, and his heterodyne receiver had a 9.5 cm collecting aperture and utilized an afc-controlled local oscillator to maintain a 70 MHz i-f. Rosner observed peak outputs ranging up to 16 dB above average levels. Differences between peak and average levels were smallest in windy weather (wind speed > 25 mph) on both clear and overcast days. The greatest differences were observed in the morning hours in clear weather. It was observed that, in the absence of rain, the behavior of the detected signal was strongly dependent on the presence or absence of sunlight; unusually constant signals were observed in the presence of high winds (see Figure 94). Rosner suggests that the sharp bursts apparent in his observations may be due to a momentary absence of turbulence-induced phase distortion in the incoming signal wave.

The heterodyne receiver used by Lucy and his colleagues at Sylvania employed spatial tracking of the angle of arrival of the signal as well as an afc system to maintain an i-f of 30 MHz. The receiving aperture was 20.3 cm in diameter. Sample probability density functions for the heterodyned signal were obtained for clear, overcast, and rainy weather conditions. The reduction in the width of the distribution (fluctuation), or the fluctuation index, as this sequence of conditions is encountered is not nearly as apparent as it is for the corresponding (simultaneous) direct detection distribution. Heterodyne sample densities were not computed for various aperture sizes; however, this work is currently being undertaken.[18]

F.E. Goodwin[20] has reported on tests performed with a 3.39μ heterodyne communication system at the Hughes Research Laboratories in Malibu, California. Utilizing a 7.6 cm collecting aperture, rms fluctuations in signal amplitude of 10 percent were reportedly observed over a 300 m propagation path.

One should keep in mind that even the smallest collection apertures used in the foregoing measurements are very large in terms of the wavelength of interest. Consequently, one can at best obtain only an estimate of the fluctuation of the average amplitude or intensity from these observations. Furthermore, one should certainly not ignore the probable influence of phase distortion in the heterodyne results. Typical results obtained by the abovementioned investigators have been gathered in Table 41, along with an indication of certain pertinent conditions attending their observations (where available).

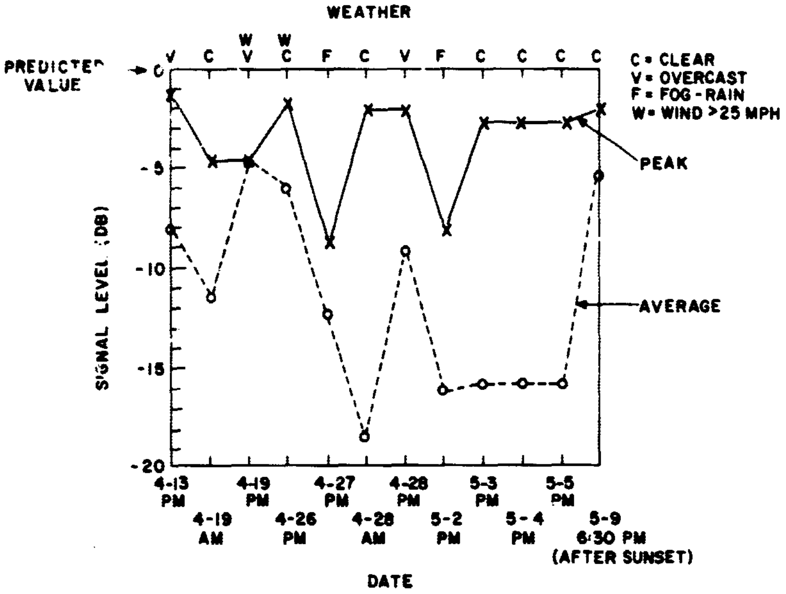


Figure 94. Plot of maximum and average signal level for 0.63μ heterodyne reception over a 2.6 km path (from Rosner[19])

Table 41

A SUMMARY OF TYPICAL FLUCTUATION MEASUREMENTS*

		Propagation Path				Receiver				
Investigator and Location	Transmitter Aperture	Path Length	Elevation Above Ground (m)	Terrain	Atmospheric Conditions	Beam Diameter	Collector Diameter (cm)	Percent Fluctuation	Peak Signals, dB Above Average	Spectral Width (Hz)
DIRECT DETECTION										
Subramanian and Collinson, BTL, Whippany, N.J.	4 cm	120 m	8	Black top pavement	Overcast, 10 mph wind	~1 cm	1-4			~70
					Heavy rain, violent wind	~1 cm	1-4			~700
	2 cm	600 m	1	Wooden walk over asphalt and gravel roof	Clear	~6 cm	15	2		
Lucy et al, Sylvania Research Lab, Waltham, Mass	1 cm (subsequently diverged)	1 km	~3	Largely paved surface	Clear	90 cm	20.3	50 σ_n % (est)		
					Overcast	90 cm	20.3	30		
					Rain	90 cm	20.3	18		
Chu, BTL, Crawford Hill, N.J.	2 cm	2.6 km	Inclined path ~15 average (est)	Varied vegetation	Clear	25 cm	5	30 average (10 - 90 range)		~200
Buck, ITSA (CRPL), Boulder, Colo.	15 cm	4 km			Clear (evening)	10 cm	2.5	63 σ_n %		~200
							5.0	53		
							11	40		
							17	17		
Peteranecz, Simons, and Taylor, Dayton, Ohio		16.5 km	70 average (est)	Varied: residential, industrial		4 m	4.3		30 max	~100

Table 41
A SUMMARY OF TYPICAL FLUCTUATION MEASUREMENTS* (Cont)

		Propagation Path				Receiver				
Investigator and Location	Transmitter Aperture	Path Length	Elevation Above Ground (m)	Terrain	Atmospheric Conditions	Beam Diameter	Collector Diameter (cm)	Percent Fluctuation	Peak Signals, dB Above Average	Spectral Width (Hz)
HETERODYNE DETECTION										
Goodwin, Hughes Research Lab., Malibu, Calif.		~300 m					7.6	10		
Lucy et al (location as above)	1 cm (subsequently diverged)	1 km	~3	Largely paved surface	Clear	90 cm	20.3	260 } σ_n % (est)		
					Overcast	90 cm	20.3	80		
					Rain	90 cm	20.3	47		
Rosner, BTL, Crawford Hill, N.J.	1.5 mm	2.6 km	Inclined path ~ 15 average (est)	Varied vegetation	Clear, windy	1.8 m	9.5		4	
					Clear	1.8 m	9.5		3 to 16 (3 after sunset)	

*All measurements were performed at 0.63 μ with the exception of Goodwin's, which were performed at 3.39 μ. All values accompanied by the abbreviation "est" were estimated from published data.
This summary is only intended to provide a convenient collection of typical results; extreme care should be used in attempting to draw conclusions from a comparison of the entries.

2.1.3 Fluctuation Rate

Fluctuation rate measurements, in the form of power spectral density analyses of the received signal, have been performed for the most part with direct detection receivers at 0.63μ. Among others, Buck,[9] Chu,[23] Hogg,[24] Lucy et al,[15] and Subramanian and Collinson[25] have contributed to these investigations. Generally, the spectra of the direct detected signals display exponential distributions at baseband with typical widths of a few hundred Hertz. The spectral shape and width have been observed to be relatively insensitive to propagation path lengths (ranging from a few meters to 145 km), or to what fraction of the transmitted beam is collected by the receiver. The spectral width, however, does seem quite sensitive to the intensity of turbulence along the path; e.g., as manifested by crosswind.

Hogg[24] first described the spectra at 0.63μ and suggested that the distribution appeared to be exponential. Subramanian and Collinson[25] were the first to suggest the invariance of the spectra to changes in path length or aperture size. These suggestions were based on data taken over 120 m and 360 m paths, in which variable apertures were employed which could collect all or only a small portion of the signal beam. Bounds on measured spectral widths ranged from 70 Hz in a calm atmosphere to 1100 Hz in rain and gusty wind. Buck has confirmed the lack of dependence of spectral shape and width on path length for large distances. He does suggest an aperture dependence for a fixed path length at frequencies below 20 Hz, however. Chu[14] has reported a slight reduction in spectral width with increasing aperture. He reports an appreciable increase in spectral width with a strong crosswind, but no corresponding increase in fluctuation power (fluctuation index). Chu has also recently described[23] the wavelength dependence of the spectral distributions. Observations at 0.63μ, 3.5μ, and 10.6μ indicate that the exponential representation is reasonably valid at each wavelength out to about 100 Hz, and that there is a definite decrease in spectral width with increasing wavelength. (Some deviation from theoretically predicted behavior is noted, however.*) The work of Lucy et al[15] strengthens the existing direct detection observations. In particular, the spectra shown for different weather conditions very graphically display the broad spectra associated with high turbulence conditions (clear, sunny weather) and with rain, and the narrow spectral width of low turbulence conditions (overcast).

Lucy and his coworkers have provided the only published[15] account of spectral measurements of the fluctuations of heterodyne-detected signals. Their work is all the more interesting in that they have obtained simultaneous direct and heterodyne detection measurements. They show that the heterodyne spectra are not grossly different from the direct-detected distributions and that the general trends with changing atmospheric conditions are similar. They remark that the heterodyne fluctuations contain stronger high-frequency components, and they attribute this to the great influence of phase distortion on heterodyne response.

From the similarity in the temporal fluctuation records, it is not surprising to find similarities between the heterodyne and the direct-detection fluctuation spectra. One should bear in mind, however, that these power spectra correspond to different moments of the signal field: the first moment in the case of the heterodyne, the second moment in the case of the direct-detection receiver. These differences do not seem to have been heretofore investigated and their examination could prove to be of considerable importance.

2.1.4 Phase Fluctuations

Buck[9] has reported observing phase fluctuations at 0.63μ over a 48.8 m path. He used a homodyne detection technique with an equal-arm Michelson interferometer in which one arm was protected from disturbances. Only a small amount of data was taken, and consequently no firm conclusions could be reached. Of the samples obtained, Buck determined that the "overall standard deviation" was 2.5 radians. From the records included in his paper, short-term fluctuations of the order of 1 radian are apparent at rates of the order of 5 Hz (which, it was suggested, were probably due to mechanical resonances).

2.1.5 Accounting for Atmospherically Induced Fluctuations

In deciding how to properly include the effects of turbulence-induced optical fluctuations, it is crucial to observe that the importance of these fluctuations is largely determined by the relative time scales involved: As indicated by the data presented above, atmospheric phenomena and the resultant optical fluctuations are characterized by time constants of the order of milliseconds. A high data rate optical communication system, however, may transmit at megabits per second. In such a system, time intervals of 10^{-6} seconds (decision interval, or bit period) are of importance. Consequently, when calculating the performance of such a system; e.g., the probability of committing an error after a 10^{-6} second observation of the received signal, atmospherically induced phase and amplitude fluctuations should not enter. The signal amplitude and phase may be assumed to be constants in such calculations, random variables fixed by a particular realization of the atmosphere. Subsequent averaging over the ensemble of atmospheres may then be appropriate.

2.1.6 Degree of Coherence

The spatial dependence of the correlation within the signal field is of interest here. This dependence is frequently parametrized by a coherence distance (roughly the distance within which the correlation between separate points in the field is significant).

Goldstein et al[27] have performed measurements at 0.63 μ and at 1.15 μ at the Air Force Electro-Optic Surveillance Site in New Mexico. They employed a heterodyne receiver with a variable collection aperture and observed propagation effects over 4 and 24 km paths. However, no definitive data were obtained at the infrared wavelength because of equipment difficulties and atmospheric absorption. The visible measurements did not encompass a sufficient range of collection aperture diameters to allow a precise determination of the coherence distance directly. Their measurements do show, however, that the coherence distance changes significantly with atmospheric conditions over relatively short time intervals (15 min). It was also quite apparent that the coherence distances were considerably smaller for the 24 km path as compared to the 4 km path. These data confirm qualitatively what one would expect on purely physical grounds. Measured values of the average refractive index structure constant (C_n) may be combined with Equation (8) of Reference 27 (an expression due to Tatarski[26]) to estimate the coherence diameter. The average value of C_n observed on the 4 km path indicates an average coherence diameter at the receiver of 1.15 cm. Comparable observations of C_n for the 4 and 24 km paths yield coherence diameters of 3.5 and 2.75 cm, respectively.

Lucy and his group, in presenting their results,[15] do not employ the notion of coherence. However, they do provide measured "heterodyne efficiencies," from which one can estimate the diameter of the coherence area. The data taken over their 1 km path at 0.63 μ indicate coherence areas with diameters ranging from 0.36 to 0.7 cm, corresponding to clear, sunny conditions for the smaller value and overcast or light rain conditions for the larger. These measurements were made with the 20.3 cm aperture. The efficiencies observed with a series of smaller apertures indicate a coherence diameter of the order of 1 cm. The angular alignment tolerance they observed can also be related to the coherence area (Section 6.4), and these observations provide an independent and consistent indication of the 1 cm value. Samples of the foregoing estimates, along with certain theoretical estimates to be discussed in Section 2.2, have been collected in Table 42.

Hinchman and Buck[28] attempted to measure intensity fluctuation correlations in a 0.63 μ laser beam at a distance of 15 km. Intensity correlation could not be detected for a photomultiplier separation of 15 cm.

In comparing his theoretical work with some observations, Reiger[29] presents some data obtained by Protheroe and Chen which indicate an intensity correlation distance of 4 or 5 cm associated with intensity fluctuations in a field of starlight. These numbers correspond to intensity correlation diameters of 8 to 10 cm.

2.1.7 Lack of Experimental Data

From the above summary, and particularly from Table 41, one can see that, although some experimental results have been obtained, the existing empirical knowledge is far from complete. More definitive and systematic experiments related to all aspects of the atmospheric optical propagation problem are needed. Of notable significance is the absence of optical propagation measurements over vertical paths; such measurements are crucial to an evaluation of an Earth-based receiver operating in a deep space optical communication system.

2.2 Theoretical Considerations

Some of the analytical work which relates to the experimental results reviewed in the previous section will be summarized below.

2.2.1 Amplitude-Related Effects

2.2.1.1 Attenuation. Predicted values of clear air attenuation have been given in an RCA report[11] and are stated in terms of percent transmission over a vertical path. For a 100 km vertical path, transmission in the visible range under "exceptionally clear" conditions is indicated to be near 90 percent; transmissions for "clear day" conditions are specified at 20 percent.

A convenient tabulation of the dependence of atmospheric transmission on wavelength and zenith angle under clear conditions may be found in the Perkin-Elmer report (Reference 4, Table 3-1), along with a brief discussion of some of the factors responsible for clear air attenuation. Elterman[34] has described a model, based on a clear standard atmosphere which includes an aerosol component, from which clear air attenuations may be determined.

Expressions for the attenuation from scattering of the mean or coherent wave in a homogeneously turbulent atmosphere have been obtained by J. B. Keller[30] and L. S. Taylor.[31] For an inner scale size of 10 cm and for refractive index fluctuations* (σ_n^2) ranging from 10^{-16} to 10^{-14}, the attenuation has been computed for the wavelengths of interest in Table 43. The values in the table do not indicate the attenuation that will be experienced by the total wave (coherent and incoherent waves), and consequently cannot be directly compared with observed attenuations; e.g., the 0.2 dB/km average level observed by Buck.[9] However, the indicated attenuations in the visible are consistent in order of magnitude with certain of Rosner's[19] heterodyne measurements.

*The appearance of $\sigma_n{}^2$ or $C_n{}^2$ depends on the choice of representation of atmospheric effects. Those qualities are closely related; see Reference 27 or Equation 6.2 of Reference 39.

Table 42

ESTIMATES OF COHERENCE DIAMETER FOR VISIBLE WAVELENGTHS

Investigator	Reference	Wavelength (μ)	Path Length (km)	Elevation (m)	Atmospheric Conditions	Estimated Coherence Diameter (cm)	
						Experimental	Theoretical
Lucy et al, Sylvania Research Lab., Waltham, Mass.	15	0.63	1	~ 3 m	clear	0.36	
					overcast	0.7	
					rain	0.72	
Goldstein, et al, AF Electro-Optic Surveillance Site, Cloudcroft, N.M.	27	0.63	4 (2 km one way)	~ 30 m (average)	$C_n = 1.1 \times 10^{-7}\ m^{-1/3}$	1.15	
					$C_n = 3.5 \times 10^{-8}\ m^{-1/3}$	3.5	
		0.63	24 (12 km one way)	~ 80 m (average)	$C_n = 1.8 \times 10^{-8}\ m^{-1/3}$	2.75	
Hufnagel and Stanley	4, 35	0.5	Vertical path through entire atmosphere		clear		~ 2
D. L. Fried	39 (Fig. 9)	0.63	1		$C_n = 1.7 \times 10^{-8}\ m^{-1/3}$		20
					$C_n = 1.7 \times 10^{-7}\ m^{-1/3}$		1.5
		0.5	1		$C_n = 10^{-7}\ m^{-1/3}$		2

Table 43

ATTENUATION DUE TO SCATTERING BY HOMOGENEOUS ATMOSPHERIC TURBULENCE

σ_n^2	10^{-16}	10^{-14}
λ	Attenuation: dB/meter	
0.5μ	1.37×10^{-2}	1.37
1.0μ	3.42×10^{-3}	3.42×10^{-1}
10.0μ	3.42×10^{-6}	3.42×10^{-3}

Attenuation from precipitation has also been examined. The RCA report mentioned above indicates only about 1 percent transmission (20 dB loss) over a 100 km vertical path through a light fog.

The experimental work of T. S. Chu[10] described earlier was complemented by a theoretical analysis of scattering and absorption resulting from precipitation in the lower atmosphere. A precise prediction of precipitation attenuation is hampered by the uncertainty in particle size distributions and forward-scattering contributions. Chu's analysis is based on the work of Van de Hulst,[32] and the results are summarized in the curves of Figures 95 and 96. The experimental data shown in Figure 93 are in good quantitative agreement with the predicted attenuations for rain, and confirm qualitatively the wavelength dependence for fog. Although 10 μ attenuation by fog is much smaller than at visible wavelengths, 10 μ attenuation may nevertheless be as great as 40 dB/km for liquid water concentrations of 0.1 gm/m^3.

2.2.1.2 Fluctuations. Various techniques for determining amplitude fluctuations have been developed. The wave optics method has been used by Tatarski,[26] Chernov,[41] and others, and P. Beckmann[33] has employed a simplified geometrical optics scheme. The analytical expressions for the amplitude fluctuations readily lend themselves to central limit theorem arguments, which indicate that a wave need penetrate a turbulent medium only a relatively short distance before its amplitude tends to obey a log normal distribution. (See Section 2.1 for comments on the experimental verification of the log normal distribution by Fried et al,[17] and also by Höhn.[21])

2.2.2 Phase-Related Effects

2.2.2.1 Phase fluctuations. Beckmann[33] has obtained an expression for the spatial dependence of phase fluctuations in a turbulent medium, as well as expressions for mean square fluctuation rates under various conditions. The fluctuation and its rms derivative are shown to vary directly with the product σ_n k L $^{1/2}$, where σ_n is the standard deviation of the refractive index, k is the wave number in a nonrandom medium, and L is the length of the propagation path. The dominant source of temporal phase fluctuations is observed to be crosswind; i.e., wind blowing transverse to the nominal beam direction. This result assumes that the temporal variation of refractive index is due primarily to drift (rather than to turbulence or diffusion).

2.2.2.2 Angle of arrival fluctuations. Beckmann also obtained expressions for the angle of arrival and has shown that the distribution of these fluctuations will be Gaussian. Hufnagel[4] has estimated the bounds on the angle of arrival fluctuations expected at a ground receiver operating at 0.63 μ. These estimates are shown in Table 44 and indicate bounds on the fluctuation expected when an initially uniform plane wave traverses the entire atmosphere.

Table 44

EXPECTED FLUCTUATION IN ANGLE OF ARRIVAL, AFTER HUFNAGEL (Perkin-Elmer Report, Table 3-3)

Zenith Angle (degrees)	rms Fluctuation: μ rad Receiving Aperture Diameter 0.131m	 1.31m
0	8.5	3.68
30	9.12	3.98
45	10.1	4.41
60	11.88	5.24

From the discussion of angular misalignment in a heterodyne receiver given in Section 6.3 and in Appendix 6, one finds that a 0.5 μ heterodyne operating with a 13 cm coherence diameter (corresponding to a coherent 0.13 m aperture) cannot tolerate angular deviations in excess of 3.85 μ rad. Similarly, with a 1.3 m coherence diameter (1.3 m coherent aperture), deviations greater than 0.38 μ rad are excessive. Under the best conditions (pointing to the zenith), Hufnagel's computations indicate possible fluctuations of the order of 2 and 10 times these values, respectively. These results indicate the need for an angle-tracking heterodyne receiver;* if the fluctuations indicated are observed, the usual "fixed direction" heterodyne would perform very poorly. Reiger[29] has also determined values for the rms angle of arrival in the visible region for various model atmospheres, and his computed values fall well within the bounds obtained by Hufnagel.

2.2.3 Degree of Coherence

A number of formulations have been put forth for the determination of the coherence properties of a wave propagating in a random medium. One of the earlier solutions is due to Hufnagel and Stanley,[35] who determined the mutual coherence function for a wave which has traversed the entire atmosphere along a vertical path. The method relies on the use of a model atmosphere, and the ground level mutual coherence function at 0.63 μ has been computed; a coherence distance slightly greater than 1 cm is obtained.[4] Similar computations at 10 μ may be performed, and one finds ground level coherence distances of the order of 30 cm.†

*Recently reported[15] rms tracking accuracies of $\pm 25\mu$ rad must be improved if fluctuations in angle of arrival are to be usefully compensated for in fields (in the 0.5 to 1.0μ region) having coherence distances greater than a few centimeters.

†With the development of an improved atmospheric model, indications are that these values are pessimistic.[38]

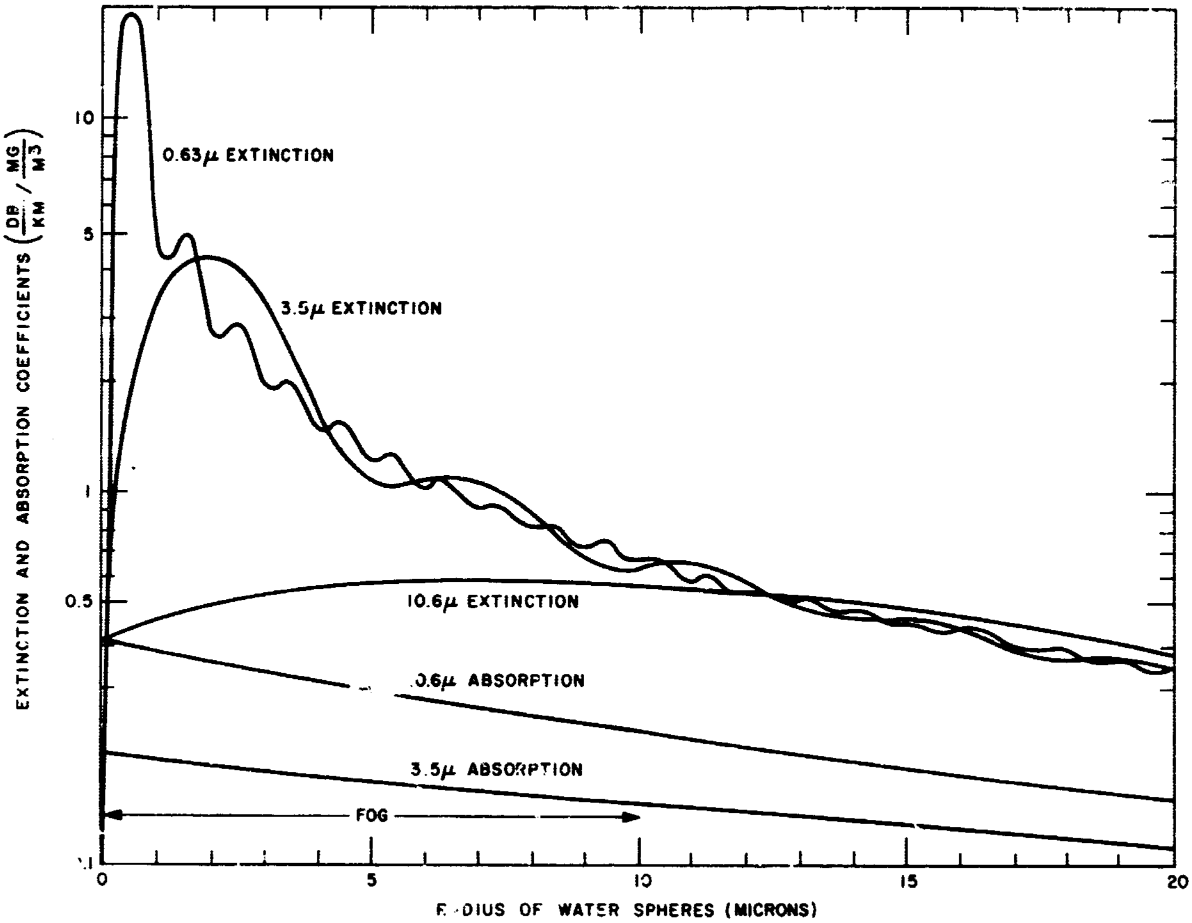


Figure 95. Theoretical extinctic and absorption coefficients (from Chu[10])

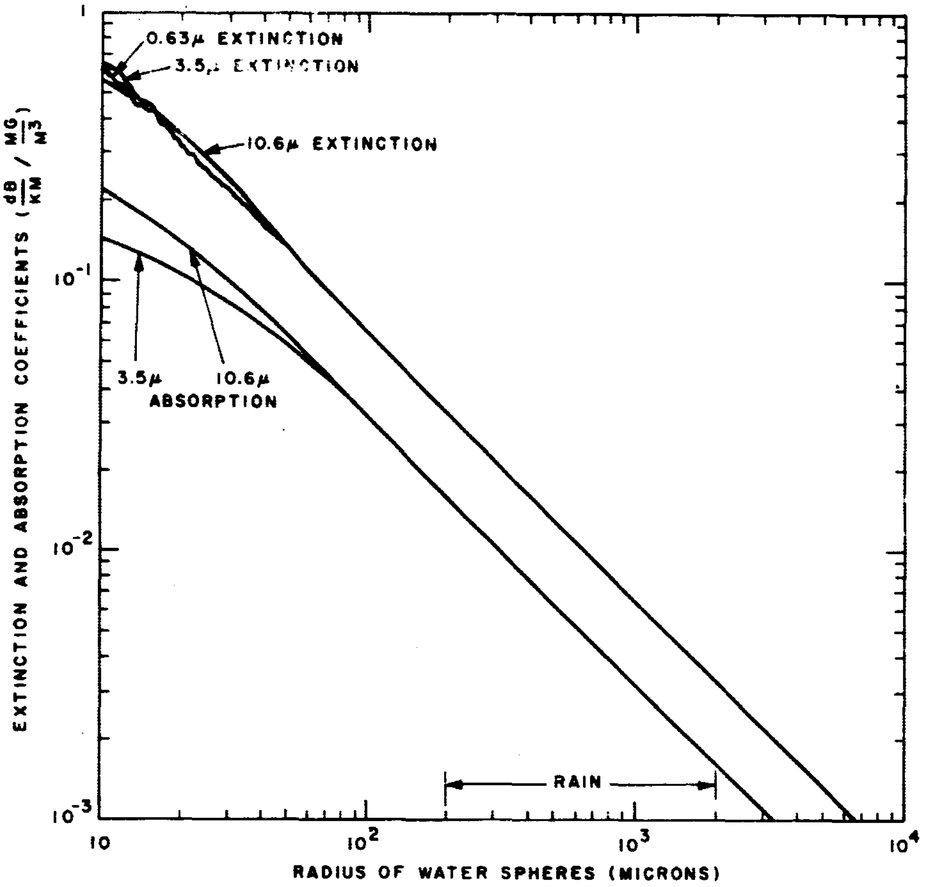


Figure 96. Theoretical extinction and absorption coefficients (from Chu[10])

L.S. Taylor[31] has calculated the mutual coherence function by means of a stochastic perturbation technique developed by J.B. Keller.[30] Taylor indicates that his analysis points up defects in certain other methods and he attempts to clarify the region where his and other existing solutions are valid.

Beran[36] has obtained a solution for the mutual coherence function in a turbulent half-space by an iterative technique. Brown[37] has employed a selective summation procedure to determine the first two moments of a field which has propagated a large distance through a weakly inhomogeneous medium. Brown has shown that his solution reduces to that found by Hufnagel and Stanley.[35]

Reiger[29] has determined the intensity correlation function for visible light transmitted vertically through the earth's atmosphere by a geometrical optics procedure. Tatarski has obtained a similar solution using wave optics techniques.[26]

Hufnagel and Stanley's work, and also Reiger's, suggests that much of the degradation from turbulence occurs in the atmospheric layer near the earth's surface. Fried[39] has computed the improvement to be expected by elevating the receiver and shows that, at 1μ, the coherence diameter can be nearly doubled by raising the receiver 3 km above ground level (a change from roughly 5 to 9 cm). These numbers correspond to zenith viewing under daytime conditions and would increase if a more optimistic atmospheric model were to be used in their determination.

The various solutions for the mutual coherence function in a turbulent medium have been obtained through use of a variety of techniques and approximations. These solutions do not yield grossly different results. The major point at issue seems to be the range of validity of the existing solutions; this is currently a controversial topic and some relevant discussion may be found in Taylor's paper,[31] and also in an article by Fried.[40]

2.3 Influence of Propagation Effects on Communication System Performance

Having observed (Section 2.1) the great difference in time scales for turbulence-related atmospheric phenomena and for high-data-rate communication systems, one may assess the effects of atmospheric-induced phase and amplitude fluctuations on such systems by considering the effects of slow, random changes in phase and amplitude on receiver performance. The effects of spatial coherence on the various types of optical receivers must also be considered.

2.3.1 Amplitude Fluctuations[55]

In the case of amplitude fluctuations which are slow compared to the communication rate, the average error probability P_e of a digital communication system may be obtained by calculating the error probability for a constant signal amplitude R, and then averaging over the distribution of amplitudes p(R).

$$\bar{P}_e = \int_0^\infty dR\ p(R)\ P_e(R) \qquad (12)$$

Consider a binary orthogonal modulation system (e.g., polarization-shift-keying) in which envelope detection is employed. For constant signal amplitude R in the presence of additive Gaussian noise with variance N, the error probability is given by

$$P_e(R) = 1/2 \exp(-R^2/4N) \qquad (13)$$

Although Equation (13) is not exact for many of the modulation systems of interest in this study, it is representative of the essential exponential dependence of error probability on signal-to-noise ratio.

There is even less basis for the choice of the fading distribution p(R) Scattering theories[41] which relate log R to a path integral suggest that R has a log normal distribution. An alternate description, which seems to have as much theoretical basis (for example, it is consistent with single scatter theories) and yet is more manageable analytically, is to assume that the received signal consists of the sum of a sinusoid with amplitude A and a narrow-band Gaussian process with variance σ^2. The resulting envelope then has the Rician distribution

$$p(R) = \frac{R}{\sigma^2} I_o\left(\frac{AR}{\sigma^2}\right) \exp -\left(\frac{R^2 + A^2}{2}\right) \qquad (14)$$

The average signal power, S, is given by

$$S = \frac{1}{2} \int_0^\infty dR\ R^2\ p(R) = \frac{A^2}{2} + \sigma^2 \qquad (15)$$

Let

$$\left.\begin{aligned} \sigma^2 &= a\ S \\ \frac{A^2}{2} &= (1-a)\ S \end{aligned}\right. \qquad (16)$$

Therefore a is the fraction of the signal power which is in the "fading mode"; $a = 0$ corresponds to a steady signal; $a = 1$ corresponds to a Rayleigh fading signal.

It is readily shown from the above that the average error probability is given by

$$\bar{P}_e = (2 + a\frac{S}{N})^{-1} \exp -\left[\frac{(1-a)\ S/N}{2+a\ \ S/N}\right] \qquad (17)$$

Note that for $a \neq o$, $\bar{P}_e$ decreases as $(S/N)^{-1}$ rather than exponentially as in the nonfading ($a = o$) case.

In Figure 97, $\bar{P}_e$ is shown as a function of S/N for $a = 0, 0.1, 0.5, 1$. It is seen that even for a as small as 0.1, 6 dB more average signal-to-noise ratio is required to achieve $P_e = 10^{-4}$ than in the nonfading case.

Diversity techniques may be employed to reduce the effects of fading. Direct detection systems which employ aperture sizes greater than the coherence area and relatively wide fields of view may achieve some of the benefits of space and angle diversity, but there is inadequate information on how amplitude statistics depend on aperture size and field of view (see Section 2.4).* Also, coding techniques (e.g., time interleaving) may be used to obtain the benefits of time diversity, with, however, some disadvantage in transmitter complexity.

Although existing information does not permit precise evaluation of the performance of optical digital communication systems over atmospheric paths, it appears clear that approaches which simply account for the atmosphere by an average attenuation may give grossly optimistic results.

2.3.2 Phase Fluctuations

The same mechanisms which give rise to amplitude fluctuations also give rise to temporal phase fluctuations in the received signal. A direct-detection optical communication system will be insensitive to such fluctuations, but they will, in general, affect the performance of a heterodyne receiver. However, if the phase fluctuations are slow compared to modulation rates of interest,† then they may be tracked in the same way that the Doppler variation is tracked. The phase fluctuations increase the minimum bandwidth of the frequency tracking circuits; but if the resulting bandwidth is small compared to the communication bandwidth, then the power required in the carrier need be only a small fraction of the total signal power. Thus, slow phase fluctuations impose no inherent penalty on communication performance.

2.3.3 Coherence Degradation

The relative importance of degraded spatial coherence within the signal field differs between direct detection and heterodyne-type receiving systems. When the collecting aperture is small and coincides with the photosensitive surface, the direct detection system is not critically dependent upon a large coherence area; such a system may perform well in a signal field with very poor spatial coherence properties. The information rate and error probability are more seriously affected by a widening of the field of view, and temporal fluctuations in signal and background fields induced by atmospheric turbulence.

In practice, however, direct detection collection apertures must frequently be made large, and a mirror or lens is required to focus the collected energy on a small photosurface. In such situations, degradation in focusing due to poor spatial coherence can become a serious problem.

Whatever configuration is employed, the performance of a heterodyne receiver is strongly dependent on the coherence areas of both signal and local fields. As shown in Section 6.3, the signal power at i-f may often be proportional to the smaller of the local and signal field coherence areas. For a propagation path traversing the entire atmosphere and terminating at ground level, Hufnagel and Stanley's calculation[4] strongly indicates that the signal coherence area will be the limiting factor at optical wavelengths. To improve the information rate for a given error probability (Section 6.4), one would like to obtain as large a signal coherence area as is possible. This improvement can perhaps most easily be accomplished by placing the receiver above as much of the turbulent ground layer as possible.

When small coherence areas prevent efficient use of a collection aperture of desired or specified size, a diversity-combining system may have to be employed. One may think of this as a technique which artificially expands the effective coherence area.

Fried[43] has recently investigated the fluctuation to be expected in the power output of a heterodyne receiver. (He computes what one might call a power fluctuation index.) He shows that when the collection aperture diameter exceeds a certain parameter r_o (essentially the coherence diameter), the fluctuation grows rapidly. However, Fried's results must be interpreted carefully because his averaging is taken over the ensemble of atmospheres. His results do not adequately relate to the power fluctuation that would be observed within a decision interval in a high data rate system, and hence do not necessarily imply that a high probability of error should be associated with heterodyne receivers.

Heidbreder[44] has shown that the form of the structure function employed by Fried is very close to that corresponding to a randomly tilting incoming plane wave. In attempting to explain a result which runs counter to what one might expect intuitively, Fried has observed the latter similarity, but he does not suggest that the predicted gross fluctuations might be significantly reduced by an angle tracking heterodyne.

2.4 Suggested Experiments

Knowledge of many physical parameters describing the turbulent atmosphere is needed. In particular, to make

*The diversity improvement resulting from large aperture direct detectors has been estimated in a recent paper by W.N. Peters and R.J. Arguello.[42]

†As noted in the next section, it is the spatial rather than the temporal character of the phase fluctuations which provide the principal difficulty.

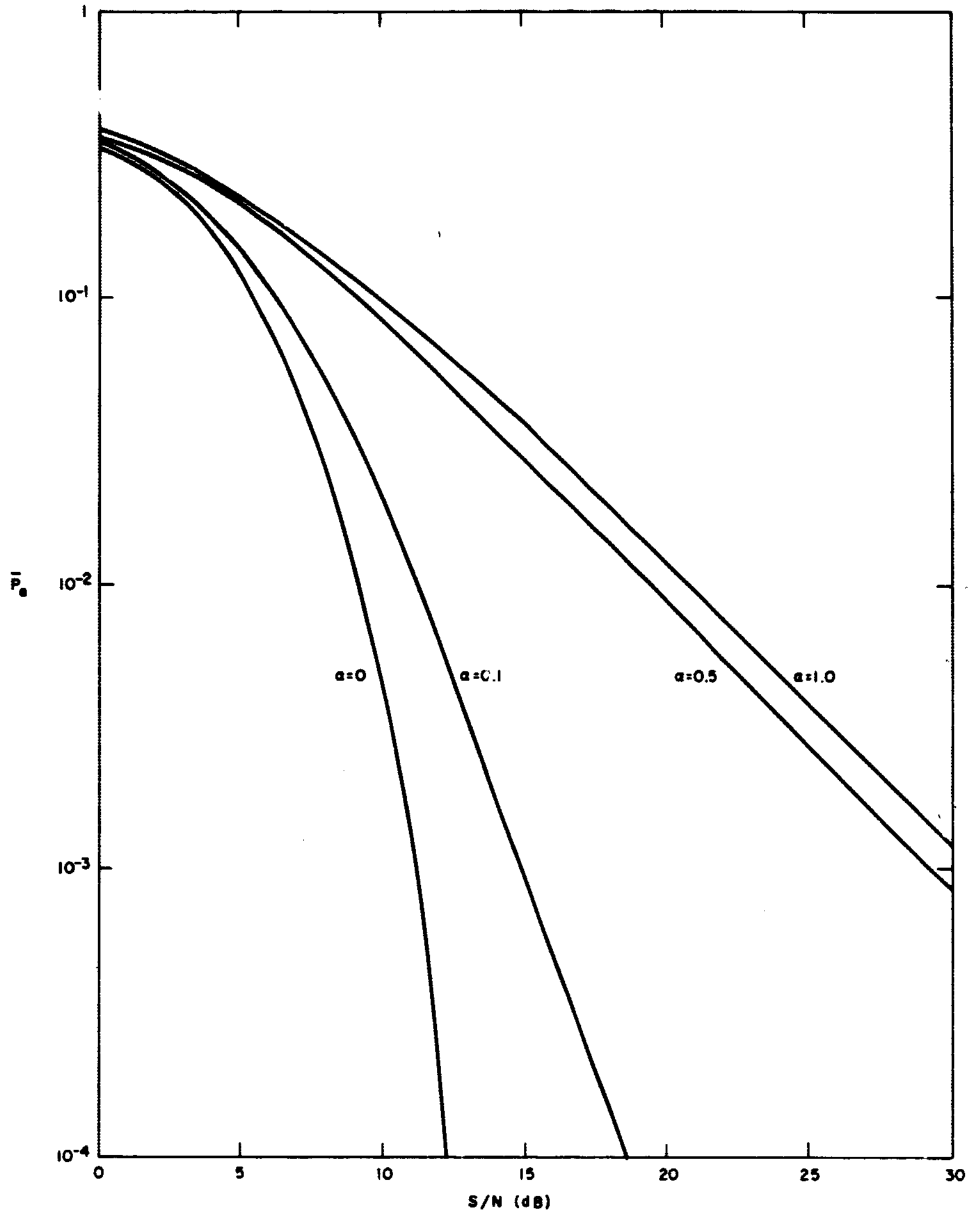


Figure 97. Effects of amplitude fluctuations on the performance of digital communication systems

meaningful use of theoretical results, reasonably accurate estimates of refractive index structure constants and structure functions – and their parametric dependence – must be obtained. Here, however, consideration will be restricted to direct field measurements needed to predict optical communication system performance. The measurements needed fall into two general categories: (1) those of pertinence to both direct detection and heterodyne systems and (2) those primarily of importance to heterodyne systems.

Type 1 Measurements

Structure and general parametric dependence of signal fluctuations; i.e., dependence on wavelength, propagation path length, size of collecting aperture, atmospheric conditions, etc. The existing fluctuation measurements have been made with narrow beams and it is known that beam wandering contributes to these fluctuations.[9,12,28,33] Perturbations internal to the beam are felt to be the dominant effect, but if possible these effects should be separated.

Type 2 Measurements

a. A general parametric investigation of the coherence distance at the terminus of a vertical propagation path is needed. In particular, the dependence on receiver height above ground level should be determined.

b. A careful study of the fluctuation in angle of arrival is necessary. A parametric investigation is again called for and particular attention should be given to the determination of whether or not the observed fluctuations be within the capabilities of angle tracking heterodynes.

Finally, it is noted that the Woods Hole Summer Study Group[45] has described measurement techniques that might prove useful in this work. The suggested experiments could well involve the use of balloon, aircraft, and satellite-borne equipment, as well as controlled laboratory experiments.

3. BEAM-POINTING CONTROL AND ACQUISITION PROCEDURES

In space communication using optical frequencies, beam pointing is of critical importance. The basic requirements on the beam-pointing control system are the ability (1) to determine the difference between the actual and the desired transmitter positions, (2) to respond to this difference by driving the transmitter to the desired position, and (3) to stabilize the transmitter at the desired position despite disturbances. The particular method of implementing these requirements depends on the required pointing accuracy. Systems with pointing accuracies of 10^{-5} radians are well within the capability of the present attitude control art. Attitude control devices and star trackers accurate to within $5(10)^{-6}$ radians are available. Figure 98 (Figure 1 of Reference 46) summarizes the pointing accuracies achievable by various attitude control devices. The performance of the Stratoscope II tracking system* is included in Figure 5-6 for comparison. For an accuracy of 10^{-5} radians, beam pointing can be achieved in an open loop mode. The space vehicle reference frame can be determined by star trackers, and the desired positions of the beam transmitter and the space vehicle can be obtained from tracking information and stored in an on-board guidance and control computer. The beam-pointing control system and the associated attitude control system acting in a closed loop mode can, by reference to the stored positions, stabilize both the space vehicle and the beam transmitter near their respective desired positions. The attainment of pointing accuracies of 10^{-6} to 10^{-7} radians will require the use of transmitting telescope optics as an attitude sensor. The telescope measurements will first be used to establish the apparent line of sight between the Earth station and the telescope, and then used for tracking the Earth station. For these accuracies, beam pointing can still be achieved in an open loop mode. However, to aid the space vehicle in the acquisition and tracking of the Earth station, a beacon at the Earth station will be necessary. For a pointing accuracy better than 10^{-7} radians, attitude sensors corresponding to a spaceborne telescope with a primary lens diameter considerably larger than 1 meter will be necessary. Unless technological breakthroughs result in a dramatic reduction of weight of the attitude sensors accurate to within fractions of 10^{-7} radians, systems with accuracy requirements of this order of magnitude will not be feasible in the near future,

In the following sections, the discussions on beam control systems and the associated attitude control systems will be confined to systems utilizing an Earth beacon.†

3.1 Space Vehicle Beam Pointing Control Systems

In addition to the Earth beacon, it is assumed that only space-to-Earth optical communications capability is required, but a low rate channel is available both from the space vehicle to Earth, and vice versa. The low rate channel will enable the space vehicle to receive Earth commands, and relay infonnation concerning its state to the Earth prior to optical commnnications. It will also serve as a back-up channel.

The expected position of the space vehicle relative to the Earth station is known to within a certain accuracy from trajectory computations made prior to launch and from trajectory measurements performed early in the flight

*Estimates of the performance of the Stratoscope II tracking system are based on a private communication from Prof. M. Schwarzchild of Princton University.

†The term Earth beacon will be used either to refer to a beacon on the Earth or on an Earth satellite.

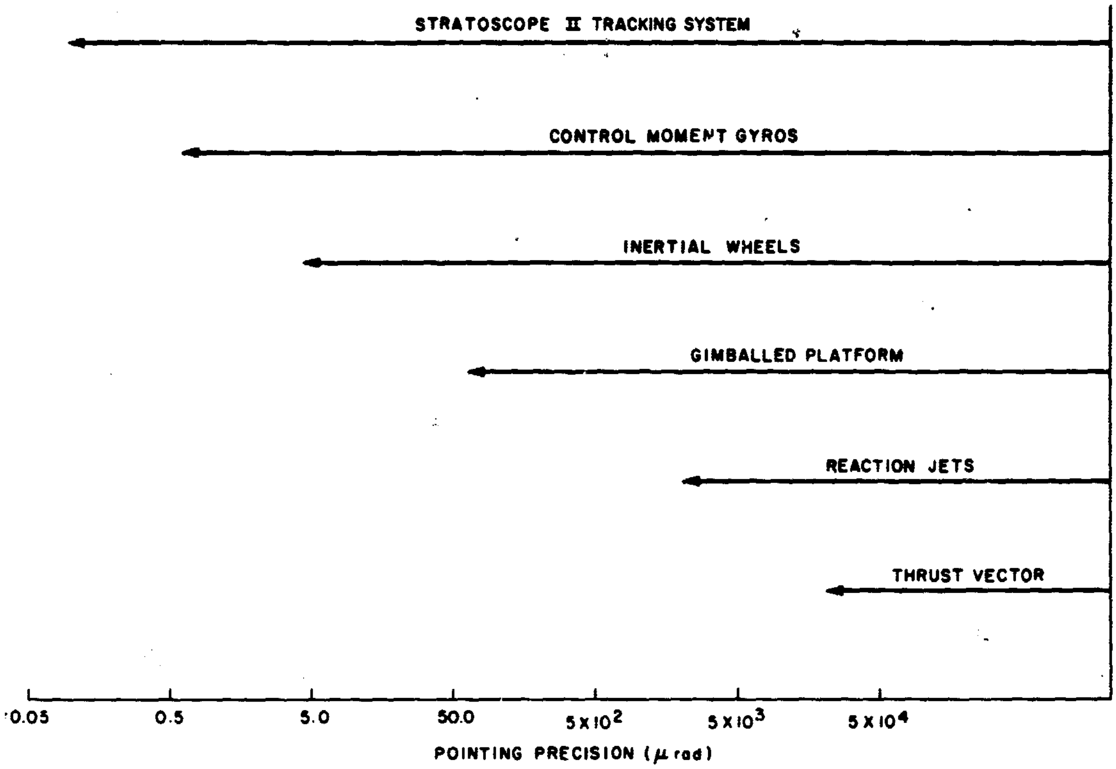


Figure 98. Point accuracies of various altitude control devices

of the space vehicle. From these data the Earth beacon can be pointed toward the expected position of the space vehicle with sufficiently wide beamwidth and intensity to assure detection and tracking by the receiving telescope on the space vehicle (see Section 4.1).

The tracking telescope on the space vehicle will share a common optical system with the transmitting telescope such as the Perkin-Elmer optical system discussed in Section 4 and in Reference 47. This arrangement, when it is suitably designed, not only results in a weight reduction but also offers considerable advantages in tracking and in precise beam pointing.

Tracking the Earth beacon by slewing either the entire telescope system or the vehicle will provide neither the desired speed of response nor the required accuracy. A transfer lens systems, together with an appropriate steering device will be necessary to track the beacon image appearing in the telescope field of view provided that the presence of the beacon in the telescope field of view can be detected. Since these devices will be light in weight and will require little actuating power, the tracking error can be eliminated almost instantaneously. Of course, the range of movement of the transfer lens system will be limited (to conserve optical quality and for mechanical reasons). A second control loop will be needed to position the optical axis of the telescope to coincide with the beacon image, thus restoring the transfer lens to its neutral position. A third control loop (the space vehicle attitude control system) can then be used to restore the telescope platform to its neutral position by re-orienting the space vehicle. (These three levels of control are designated by the letters C, B, and A respectively in Section 3.3.)

For the last two control loops, the actuators may be a combination of control moment gyros (CMG's) and reaction jets. The CMG's will be needed to provide the necessary accuracy, while the reaction jets will be needed to desaturate the CMG's and to provide torques for high rate maneuvers. When the telescope is directly mounted on the space vehicle, the second and the third loops mentioned above can be combined into one.

The apparent line of sight between the Earth beacon and the tracking transfer lens is chosen as one of the reference directions, not simply because it appears as a natural selection, but mainly because the telescope system provides the best available attitude error sensor aboard the space vehicle. Based on the tracking error, the transfer lens movement together with the telescope and the vehicle attitude control systems hopefully will be designed to quickly reduce the tracking error to an acceptable level. For the case at hand, the acceptable tracking error should be less than the tolerable pointing error (10^{-6} to 10^{-7} radians).

To account for the long transit time delay between the vehicle and the Earth station in the case of a deep space mission, the optical axis of the transmitter transfer lens must be offset from the apparent line of sight between the tracking transfer lens and the Earth beacon. This point-ahead angle varies slowly with time as function of the relative velocity between the space vehicle and the Earth station. For a Mars fly-by, a typical point-ahead angle can be shown to be less than $2(10)^{-4}$ radians and the variation of the point-ahead angle is of the order of 10^{-7} radians/hrs, as shown in Figure 99 (Reference 48, p. 54).

The point-ahead angle can be computed from orbit information and intorduced to the beam transmitting transfer lens in a feed-forward mode. Complete feed-forward operation for an extended period of time may not be advisable because no means is available to compensate for errors due to either a long-term component drift or to inaccuracies in the point-ahead information. These can be taken out either by an additional feedback loop (the grand loop) or by a periodic correction from the Earth, based upon tracking information. Thus the optical system on the space vehicle must perform the following functions:

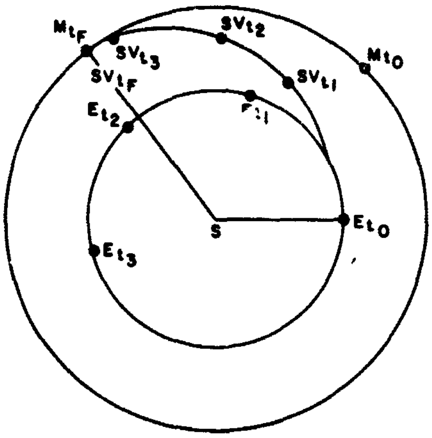
1. The development and utilization of appropriate error signals for tracking the Earth beacon
2. The generation of point-ahead signals and the utilization of these signals for pointing the Earth-bound beam.

To combine these functions in a single optical system, each with adequate freedom of movement and precison, will require much ingenuity on the part of the optical designer. The Perkin-Elmer system described in Chapter 3, Section 4 and in Reference 48 appears to be able to combine the above mentioned functions adequately.

From the above discussion a multilevel, multiloop space vehicle beam pointing control configuration is evolved. The block diagram for it is shown in Figure 100.

The quality of stability of the control configuration shown in Figure 100 depends only upon the quality stability of each individual loop (Appendix 4; see also Reference 49 for a more complete discussion). Thus, the controller design for each loop can be individually undertaken. Conventional design methods can be utilized here without any difficulty.

In Figure 100 two error sensors are shown in a parallel configuration; one has a wide-view window for developing the error signals used in aligning the tracking telescope optical axis with the earth beacon image. The other has a narrow-view window for developing error signals for the transfer lens to track the beacon image. Since the sizes of the view windows are different, it is possible to cascade the error sensors optically so that the narrow-view fine error sensor will stay inactive until the tracking error is reduced to an acceptable level. In Figure 100 movement limiters are shown to reflect the physical constraints placed upon the motion of the transfer lens and the telescope platform. A dead band is introduced between the vehicle attitude control loop and the telescope-platform gimbal-angle pick-off so



TIME	$\lvert R \rvert$ RELATIVE POSITION $\times 10^6$ Km	$\lvert \dot{R} \rvert$ RELATIVE VELOCITY Km/sec	$\lvert \alpha \rvert$ POINT AHEAD ANGLE μ rad	$\lvert \dot{\alpha} \rvert$ ANGULAR VARIATION μ rad/hr
t_1	20.6	10.3	50.4	0.131
t_2	127.6	16.0	100.5	0.05
t_3	189.2	32.5	190.5	0.078

$T_{E-M} = t_F - t_0 = 0.525$ YEAR

TRANSFER ELLIPSE: $\epsilon = 0.25, a = 1.31$ a.u. $\cong 1.96 \times 10^8$ Km

$R_E = 1$ a.u. $\cong 1.5 \times 10^8$ Km

$R_M = 1.52$ a.u. $\cong 2.3 \times 10^8$ Km

Figure 99. Typical point-ahead for Earth-to-Mars flyby

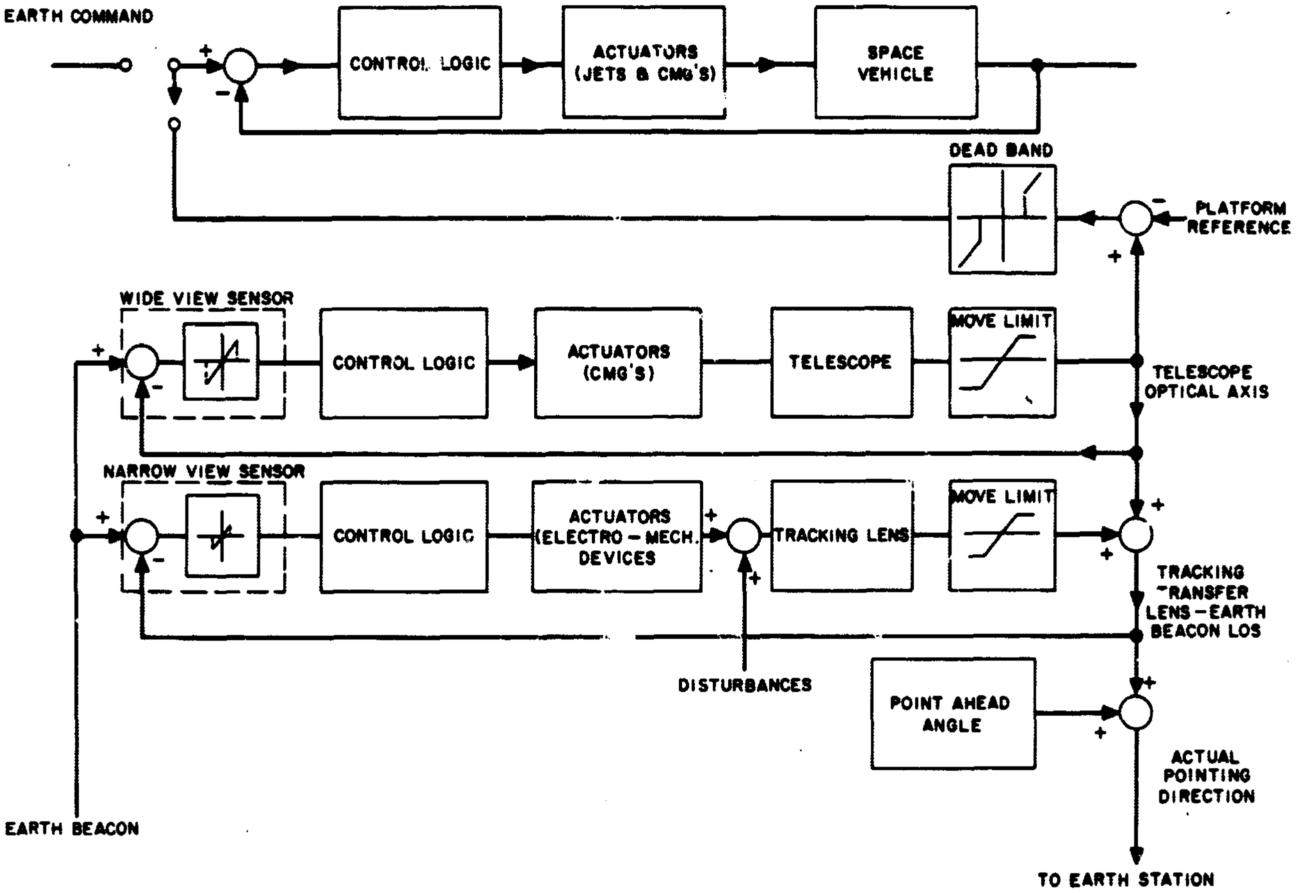


Figure 100. Space vehicle beam pointing control configuration with open loop point-ahead

that the vehicle attitude need not be adjusted until the gimbal angle becomes fairly large.

It is worth noting the following:

1. The location of the Earth station receiver has not been specified. The point-ahead computation should account for the angular difference between a ground receiver and a satellite receiver.
2, When the Earth beacon is situated on the ground, the presence of the atmosphere will affect the beamwidth and the intensity of the beacon. To assure adequate intensity of the beacon image at the space vehicle, a pulsed laser with time gating at the vehicle receiver may have to be considered (Section 4.1). The stability and control design problem may then be considered in the discrete time domain rather than the continuous time domain. This, however, does not alter the nature of the problem.
3. In Figure 100 it was assumed that the Earth beacon image will appear in the space vehicle tracking telescope field of view, and that the transmitting telescope transfer lens is pointing ahead in an open loop mode. The disturbance effect of the transmission beam (mostly atmospheric), the inaccuracies in the point-ahead signal due to misalignment and calibration errors, and the long-term component drifts are not compensated for. The acquisition of the Earth beacon and the compensation of the above-mentioned errors will be discussed in the following section.

3.2 Acquisition Procedures

Before the beam pointing control system shown in Figure 100 can track the Earth beacon, the beacon image must appear in the tracking telescope field of view. Similarly, before optical communication can commence, the Earth-bound beam must also appear in the view of the receiver. In both cases, acquisition procedures will be required to insure successful operation.

For the space-borne optics to acquire the Earth beacon signal, it is assumed that the beacon is sufficiently powerful (see Section 4.1) to provide adequate SNR ratio at the tracking telescope detector against the Earth-shine background, so that subsequent tracking of the beacon image is possible. It is assumed that the position of the Earth station with respect to the space vehicle can be determined to within some known accuracy such that, when the space-borne telescope is pointed toward the predicted position of the Earth station, the field of view of the telescope is sufficiently wide to encompass the station position uncertainty.

To accomplish Earth acquisition, the vehicle is first commanded to align one of its body axes (the Sun tracker axis) with the Sun. Coarse Sun sensors with nearly 360 degree of field of view and reaction jets provide the error signals and the actuating power for bringing the sun within the view of fine Sun sensors. Fine Sun sensors, such as the OASO critical angle prism Sun sensor which is being developed for NASA by Honeywell* are capable of resolving pointing errors to within 10^{-5} radian or less. Using these sensor outputs, CMG's align the Sun tracker axis with the Sun. The Sun line together with the apparent line of sight between the vehicle and the Earth beacon (yet to be acquired) provide the vehicle attitdue reference frame. Prior to alignment of the Sun tracker axis, the telescope will have been offset from the Sun tracker axis through platform rotation by the angle formed by the intersection of the Sun and Earth lines. This angle will have been precomputed from trajectory information and will be stored in the on-board guidance and control computer. For a Mars flyby, this angular variation is typically of the order of 1 degree/day, which implies that ample time is available for a more accurate angle estimation based on post-launch tracking data. The relatively slow rate of the angluar variation also implies a small on-board computer memory requirement. If this angle can be predicted accuractely to within 1/2 degree, then a telescope with a 1 degree field of view (such as the Perkin-Elmer system discussed in Chapter 3, Section 4) will assure Earth acquisition.

(From Mars, the Earth subtends an angle of 8×10^{-5} radians and the orbit of a satellite at synchronous attitude subtends and angle of 1.2×10^{-4} radians. Thus, if the Earth is acquired, the Farth beacon will also appear in the 1 degree field of view of the tracking telescope and beacon tracking will be possible if the beacon power is adequate.)

After the Sun tracker axis is aligned with the Sun, the vehicle is commanded to rotate slowly about this axis (e. g., 2.5 rev/hr) to acquire the Earth. When the magnitude gate and the logic circuits of an Earth shine detector indicate that the Earth has passed through the field of view of the telescope, the space vehicle will stop its rotation. The space vehicle is then commanded to rotate in the opposite direction about the Sun tracker axis at a slower rate (e.g., 0.5 rev/hr) until the Earth reappears in the telescope's field of view. Upon the reappearance of the Earth the rotation command is removed and the beam pointing control system shown in Figure 100 is actuated. The residual motion of the vehicle (0.5 rev/hr) will be sufficiently slow that the vehicle tracking telescope transfer lens will have no difficulty in tracking the Earth beacon. With the apparent line of sight between the Earth beacon and the tracking telescope transfer lens established, the point-ahead angle is added vectorially to the reference direction to give the pointing direction for the Earth bound laser beam.

It was noted previously that the angular information required for Earth acquisition is a function of the trajectory

*Based upon private communication from Jerry Kollodge of the Radiation Center at Honeywell.

time. This information when refined can also be used to generate the point-ahead angle. Taking the difference between an angle corresponding to trajectory time t and an angle corresponding to trajectory time t + T_d where (T_d is the round trip transit time delay) will give the magnitude (in the sense of polar coordinates) of the point-ahead signal. The orientation of the point-ahead signal can be referenced to a line which is the projection of the vehicle-Sun line into a plane perpendicular to the apparent line of sight between the vehicle telescope transfer lens and the Earth beacon. It follows that the point-ahead information can be stored in triplets of magnitude, orientation, and trajectory time in the guidance and control computer. The accuracy of the angular information required for the point-ahead signal must be order of magnitude better than the accuracy required for Earth acquisition and will require correspondingly larger computer memory.

If the initial pointing of the optical beam fails to achieve optical communication, a scanning procedure, to be described below, can be used to assist acquisition of the Earth-bound signal.

Assume that the Earth beacon is being tracked but, due to an equipment misalignment, the initial pointing of the laser beam fails to achieve optical communication. A scanning signal is superimposed on the nominal pointing direction. The beam width may be broadened to reduce the number of lines of scan needed to cover the field of search which is chosen to encompass the uncertainty of the Earth station location. Each scanning position in the chosen field of search is coded so that its position with respect to the nominal pointing direction is determined. The search pattern can be in the form of a square raster or a divergent spiral, or it can be performed in any sequential manner which covers the entire field of search. The pointing error is detected at the Earth station receiver when one of the coded scanning beams is received. Correction signals for the nominal pointing direction can be generated either at the Earth station or on the space vehicle. In the latter case the error code is returned to the space vehicle via the low-rate channel; the on-board guidance and control computer processes it to generate the correction signal. It follows from this discussion that, for any given field of search, the pointing error is quantized in terms of the beam size used to scan through the field of search. When the pointing error is reduced to zero, it simply implies that the location of the Earth station receiver is known to be within an area corresponding to the scanning beam size. This being so, it is obvious that the next step is to choose a field of search corresponding to this angular area and to choose a correspondingly smaller scanning beam size. It will be convenient to use a fixed search pattern so that the same codes can be used over and over again to denote the same relative position of the scanning beam with respect to the nominal pointing direction regardless of the size of the actual field of search.

Repetition of steps like this will successively reduce the beamwidth to a level suitable to commence optical communication. By nature of the process, the Earth station location is assured of being within the area covered by this beamwidth. Thus, acquisition will be achieved and biases will be reduced to an acceptable level during the process.

As an example, suppose the initial uncertainty of the location of the Earth station receiver can be confined to a $[50(10)^{-6}]^2$ steradians field of search, and a scanning beam of width $25(10)^{-6}$ radians beamwidth is chosen that the scan pattern consists of 4 positions A, B, C, and D. Then reception of a coded beam corresponding to position A will allow the location of the Earth station receiver to be confined to a $[25(10)^{-6}]^2$ steradians area. In the next step, a $12.5(10)^{-6}$ radians beam can be used to cover the $[25(10)^{-6}]^2$ steradians field of search. Reduction of the pointing error to "zero" would narrow the location of the Earth station to one of the four $[125 \times (10)^{-6}]^2$ steradians area. Successive steps similar to the above can then be used to achieve the desired accuracy.

For each fixed field of search, the space vehicle – Earth station-space vehicle loop (the grand loop) can be described by a dynamical system with a delayed input. Although the delay is truly time varying, the variation is sufficiently slow (in relation to control response time) for the delay time to be considered a constant. A controller design procedure for the error removal in the grand loop is shown in Appendix 5.

The total control response time required for reducing the quantized pointing error to "zero" for each chosen field of search can be separated into two terms, a gain independent term $\overline{T}_d$ corresponding to the round trip delay time and a gain dependent term T_c (see Figure 3, Appendix 5). If m steps are required to reduce the initial size of the scanning beam to the communications beam size and if T_e is the time required for the space-borne optics to acquire the Earth beacon, then the total acquisition time T_{ac} can be estimated by the simple expression

$$T_{ac} = m(\overline{T}_d + T_c) + T_e$$

Thus, for the example considered previously, a factor of two reduction in the scanning beam size would require nine steps to reduce the field of search from $[50(10)^{-6}]^2$ steradians to $(10^{-7})^2$ steradians. At Mars distance, typically $T_d \approx 20$ min. and $T_e \approx 60$ min. Thus, the minimum time required for acquisition is 240 min (assuming $T_c = 0$).

On the other hand, a reduction of the beam size by a factor of 10 during each step of change of the field of search requires m = 3, so that the minimum acquisition time is $T_{ac} = 120$ min. The reduction in acquisition time is achieved at the expense of more complex codes for the scanning beam.

Once acquisition is completed, the normal communications mode will begin. During this mode, because of the difference in the dimensions of the illuminated area and the dimensions of the Earth-station receiver, open-loop pointing can be realized if the pointing error is periodically corrected. On the other hand, it might be desirable to avoid the costly process of reacquisition by correcting the pointing error in a closed-loop fashion. Error sensing, in this case, can be achieved in an analog manner by recording the variation in the intensity of reception as the beam transmitter scans through a circular pattern (see Chapter 3, Section 4 for possible means of error generation).

The block diagram modifying the part of Figure 100 to accommodate the closed-loop acquisition procedure described above is shown in Figure 101.

The additional equipment required to implement the closed-loop acquisition scheme shown in Figure 101 will be the electronics for introducing the biasing signals, a mechanism for introducing different beam divergence, and additional memory for the control and guidance computer.

3.3 Attitude Control and Tracking

In Section 3.2 the hierarchy of beam-pointing control and the associated attitude control loops was outlined. Here a more detailed discussion of the three levels of control will be given.

Three levels of control will be necessary: (1) control of the vehicle as a whole to the order of one degree or better, (2) control of the main tracking and transmitting telescope and related equipment to an accuracy of better than $3(10)^{-5}$ radian, and (3) control of the transfer lens to achieve fine tracking to an accuracy of $5(10)^{-6}$ radian. These subsystems are necessary to save control energy that would otherwise be required to point the vehicle to high precision and also to quiet the tracker, isolating it from disturbance forces caused by random telescope motions. The three levels of control will be discussed in order.

Attitude control of the entire vehicle will most likely be achieved by gas jet – control moment gyro actuation and star tracker – beacon tracker sensing. It is known that with accurate sensing of angular errors a gyro – gas jet system can control attitude to the order of $5(10)^{-6}$ radian or better.

Studies of the orbiting astronomical observatory (OAO) attitude control system have shown that, even in relatively severe terrestrial torque environment, $5(10)^{-6}$ radian (rms) error can be maintained (References 46, 50, and 51). This is accomplished by using a precisely balanced air bearing supported table (vehicle) controlled by three control moment gyros. The gyros must be periodically unloaded by firing gas jets to "dump" momentum stored in the gyros by gravity unbalances and external disturbances. The experiments cited were carried out at the Boeing Company, at the General Electric Company, and at NASA Ames Research Center. At the latter laboratory, an elaborate chamber is under construction to isolate the experiment from air current and seismic disturbances; with this new facility, they expect to be able to demonstrate steady-state pointing accuracies of $5(10)^{-7}$ radian rms. These experiments indicate that the necessary vehicle control accuracy may be achieved if adequate sensing accuracy can be ensured.

The tracking telescope (level B) can be controlled by tracking the beacon signal and a star. The telescope and related equipment may be connected to the main vehicle by flexure bearings and torqued by torque motors or other low-friction electromechanical devices. Control of such a system to 10^{-5} radian is well within the capability of the servo art. Separate control of the telescope may save control energy as compared with control of the entire vehicle to 1/10 arc minute. The specific tradeoff in terms of equipment complexity, weight, reliability, and other factors must be carried out for each vehicle design.

Perkin-Elmer has designed the telescope mount for Stratoscope II using flexure pivots and electromechnaical drive mechanisms. The vehicle itself must be (vibrationally) very quiet so that large vehicle disturbances are not transmitted via the telescope mount to the transfer lens system. In this manner, Stratoscope II was able to point at a star to $5(10)^{-6}$ radian rms accuracy. For the system under consideration, the vehicle would normally be rather "noisy" from gas jet firings and moving parts. It will be necessary to go to great lengths to quiet the vehicle in order to reduce disturbances of the tracker. (See the discussion in Section 4, Chapter 3.)

The control at level C, the tracking system, is most crucial in terms of ultimate accuracy limits and thus pointing capability of the optical communication beam. A system similar to that designed by Perkin-Elmer for space application (Reference 47) seems essential. The basic elements of this tracker are (1) a transfer lens suspended by a torsion bearing and controlled in position by linear electromechanical devices and (2) an image-dissecting prism which detects the position of the image blur circle. Four detectors are used, one for each quadrant of the image-dissecting prism.

There are two main sources of error in this tracking system: (1) the shot effect and background noise caused by random photo impacts on the photomultiplier tube and (2) the random acceleration of the telescope attitude motion. An analysis of these effects is given in Section 4.2 to establish the tradeoff between reducing the effect of telescope disturbances and raising the number of photons per second (beacon signal power) that enter the receiving aperture.

The random accelerations of the telescope cause a random motion of the transfer lens. This causes the pointed laser beam to wander because common optics are used for the beacon tracker and transmitter. To decrease this disturbance, tighter control (wider servo bandwidth) must be applied in the tracking loop. However, the increased tracker bandwidth allows more of the (white) error detector noise and background noise to be amplified by the servo loop.

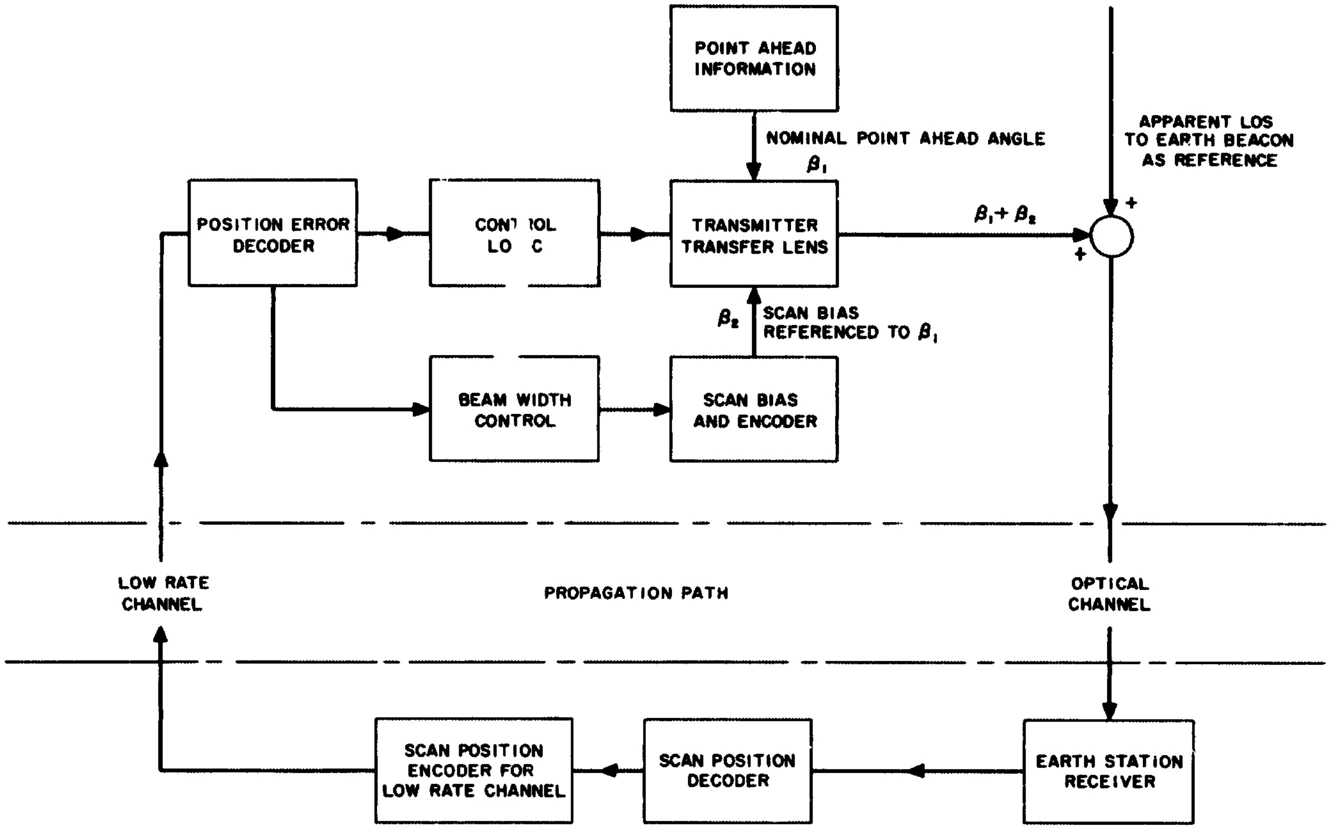


Figure 101. Modification of beam pointing control for closed loop acquisition

This also appears as a random motion of the transfer lens, producing the same effect as the acceleration noise. There is a tradeoff between good acceleration response (tighter loop) and good detection noise response (narrow-band loop). This tradeoff is the essence of the tracker design. The possibility of achieving high beacon power and thus lower effective detector noise will be analyzed later. Here it will be assumed that the detector noise level is fixed and the requirements for reducing disturbing accelerations on the transfer lens system will be emphasized.

A rather wide beacon beam may be required from the Earth station to illuminate the tracking optics. On command the tracking system aboard the spacecraft will begin a scan of its field of view. If the signal is strong enough and the disturbing accelerations low enough, the acquisition will be made. Then the vehicle will transmit a signal to the Earth station, allowing the Earth station to refine the beacon beamwidth based on the received signal direction from space.

The critical time period is the initial acquisition period of the Earth beacon signal. During this time and the time before the beacon beam can be narrowed, the vehicle must be "quieted" by closing down all possible systems that might disturb the tracker optics. In particular the spacecraft attitude control system can be turned off, leaving only the telescope controls. These measures would maximize the chances of acquisition. Specific levels of beacon power and disturbance level will determine actual feasibility.

3.4 Conclusions

In this section, the beam-pointing control system configuration, the acquisition procedure, and the physical limits to pointing a laser from deep space have been considered. It appears that the point-ahead system hardware and prediction capability allow point-ahead to within 10^{-7} radian, hence the attainment of pointing accuracy to within 10^{-6} to 10^{-7} radian depends on the ability to track the beacon to within fractions of this order of accuracy. The detailed analysis shown in Section 4 appears to show that this sets a requirement for beacon power which is very difficult to obtain with currently available devices. In addition to this difficulty, successful beacon tracking (thus beam pointing) to this order of accuracy also requires an extremely quiet vehicle, as indicated by the experience of Stratoscope II. The telescope should be tightly controlled and all friction or vibration-producing torque sources should be eliminated. Since the parameters governing the acceleration and acting on the tracking transfer lens system are not well known, further analytical studies of the attitude and tracking loops should be made to determine the effects of disturbances on tracking. Correlation of the analytical studies with tracking simulations using actual hardware should be made.

As the next step in development of a pointing system, it is recommended that a piece of developed tracking hardware, e.g., the Perkin-Elmer tracker, be mounted on a working air-bearing table to simulate all three levels of attitude control and tracking hardware. The air-bearing table, e.g., the facility at NASA-Ames Research Center, would be controlled at various levels of pointing accuracy, while the tracking system and telescope servo performed its function on a laser signal and simulated background source. The signal and background sources would be suitably attenuated to be equivalent in signal and noise properties to a beacon with Earthshine background at planetary distance. Efforts would be made to simulate actual mission conditions for each part of the system. Such an experiment will play a crucial role in determining the feasibility of a deep space laser pointing system.

4. BEACON CONSIDERATIONS

This section analyzes the power required by a cooperating optical beacon, at or near the Earth. This beacon appears to be an indispensable part of the system when very narrow communication beams, due to the use of visible frequencies, are employed. In Section 4.1, the power required for initial acquisition is examined. In this phase, the beacon beam must be sufficiently broad to assure illumination of the space vehicle; hence the signal power intensity at the space vehicle will be smaller than when a narrow beam is used. In Section 4.2, the tracking loop is analyzed, and the implications for the beacon power are noted. This analysis is a steady-state analysis, and the numerical examples chosen are based on the vehicle disturbances experienced with Stratoscope. It is possible that the level of disturbances and their frequency will be slightly lower than Stratoscope figures during the acquisition phase, as the space environment may be more benign. It might furthermore be possible to perform the tracking control (after acquisition) with a narrower beam than the acquisition; hence, the beacon power required may be somewhat less than for acquisition.

4.1 Beacon Power Requirements for Acquisition

An optical beacon from the earth may be required to point a narrow-beam optical communication signal from a deep space vehicle to an Earth receiver.* The beacon beam should be sufficiently broad to intercept the space vehicle on an open loop basis. Furthermore, in the acquisition mode, the space vehicle receiver must have a sufficiently wide field of view to ascertain that it is receiving the beacon, and the acquisition must occur in the presence of considerable background radiation (Earthshine).

*"Earth receiver" is used here to indicate either a ground-based or earth satellite receiver.

The Earth subtends a solid angle of about 10^{-8} steradian from Mars 10^8 miles away. Since, in the acquisition mode the field of view will certainly be greater than this, the acquisition receiver in the spacecraft must operate in the presence of a full Earthshine background. In the visible region this corresponds to a radiant intensity π_ν of about $6(10)^{-22}$ watts/m^2 - Hz (see Section 1). Assuming 1Å wide optical filters which correspond to a bandwidth of about 10^{11} Hz, the background noise per unit area is $6(10)^{-11}$ watts/m^2. If a 10 watt CW optical beacon with a beamwidth of 1 milliradian is assumed and no allowance is made for atmospheric attenuation, then there would be a beacon level of 10^{-15} watts/m^2 at the space vehicle. It is clear then that with a CW beacon the acquisition receiver will be strongly background-noise limited, and excessively long integration times would be required for detection. This suggests that a much narrower beamwidth and/or a high-power pulsed beacon should be employed.

If it is assumed that dark current and receiver thermal noise are negligible, it follows from the results of Section 7 that the time required for detection is given by

$$T = 30 \frac{\left[1 + \dfrac{\pi_\nu W}{\pi_s}\right]}{\dfrac{\pi_s A_R \eta\tau_o}{h\nu}} \tag{18}$$

where π_s = received signal power per unit area

A_R = area of receiving aperture

η = detector quantum efficiency

τ_o = transmissivity of optical receiving system

π_ν = radiant intensity from Earthshine = $6(10)^{-22}$ watts/m^2 - Hz

W = optical filter bandwidth.

It is assumed in Equation (18) that a signal-to-noise ratio of 15 dB suffices for acquisition. This corresponds to the calculation of the required signal-to-noise ratio for tracking control (see succeeding section), and approximates the 13 dB figure which can be used for binary communication (Section 7). In practice, the signal-to-noise ratio required for acquisition may be as high as 20 dB, as suggested by Stratoscope II experience (see Chapter 3).

The signal power per unit area at the receiver is approximately related to the transmitter power and beamwidth by

$$\pi_s = \frac{1}{2}\tau_A \frac{P_t}{(R_o\theta)^2} \tag{19}$$

where P_t = transmitter power

τ_A = transmissivity of optical transmitter and path from transmitter to receiver

R_o = range

θ = beamwidth.

As an example, consider a 1 meter diameter receiving aperture, $\eta\tau_o = 0.1$, $\pi_\nu = 6(10)^{-22}$ watts/m^2 - Hz, W = 10^{11} Hz, τ_a = 1, $R_o = 10^{11}$ meters, and λ = 0.5 microns. With these parameter values, Equations (18) and (19) are used to calculate transmitter power for beamwidths of 10^{-4} and 10^{-5} radians, and the results are shown in Figure 102. For the longer detection times, the curves correspond to the case where the background noise in the optical filter bandwidth greatly exceeds the signal. In this regime P_t is given approximately by

$$P_t \approx \sqrt{\frac{30\pi_\nu W h\nu A_R}{\eta\tau_o T}}\;\frac{2}{\tau_A}\;\frac{(R_o\theta)^2}{A_R} \tag{20}$$

The detection time must be much less than the period of the principal angular disturbances of the telescope system. Stratoscope II experience suggests that these disturbances may be as high as 10 Hz and that T must be less than 10^{-2} seconds (see Chapter 4 and the following section). It follows then from Figure 102 that a CW beacon having a beamwidth of 10^{-4} radian requires several hundred watts. This is beyond the limits of existing CW lasers in the visible region.

It follows from Equation (20) that P_t is proportional to the square of the beamwidth, so that reasonable beacon powers may be obtained if the beamwidth can be narrowed below 10^{-4} radians. Although the tracking system has an angular noise well below 10^{-4} radians (see Chapter 7) it is questionable whether biases can be kept this small.

It is possible, of course, to acquire at shorter range and to narrow the beacon beam as the range increases. However, there is the need for reacquisition following handover, and the question again arises as to the narrowest beam which can be pointed from the ground station, on the basis of past trajectory information, with full confidence of illuminating the space vehicle. The use of $\theta = 10^{-4}$ radian for this number appears optimistic particularly if the ground station is on earth and atmospheric degradations are encountered. (In that case, there is a lower bound on the beacon beamwidth caused by atmospheric spreading.)

A reduction in beacon power may be achieved by reducing the optical filter bandwidth (1Å was assumed above) or the post-detection filter bandwidth (1/T = 160 Hz is required in the analysis of the tracking loop operation in the following section), but P_t varies only as the square root of

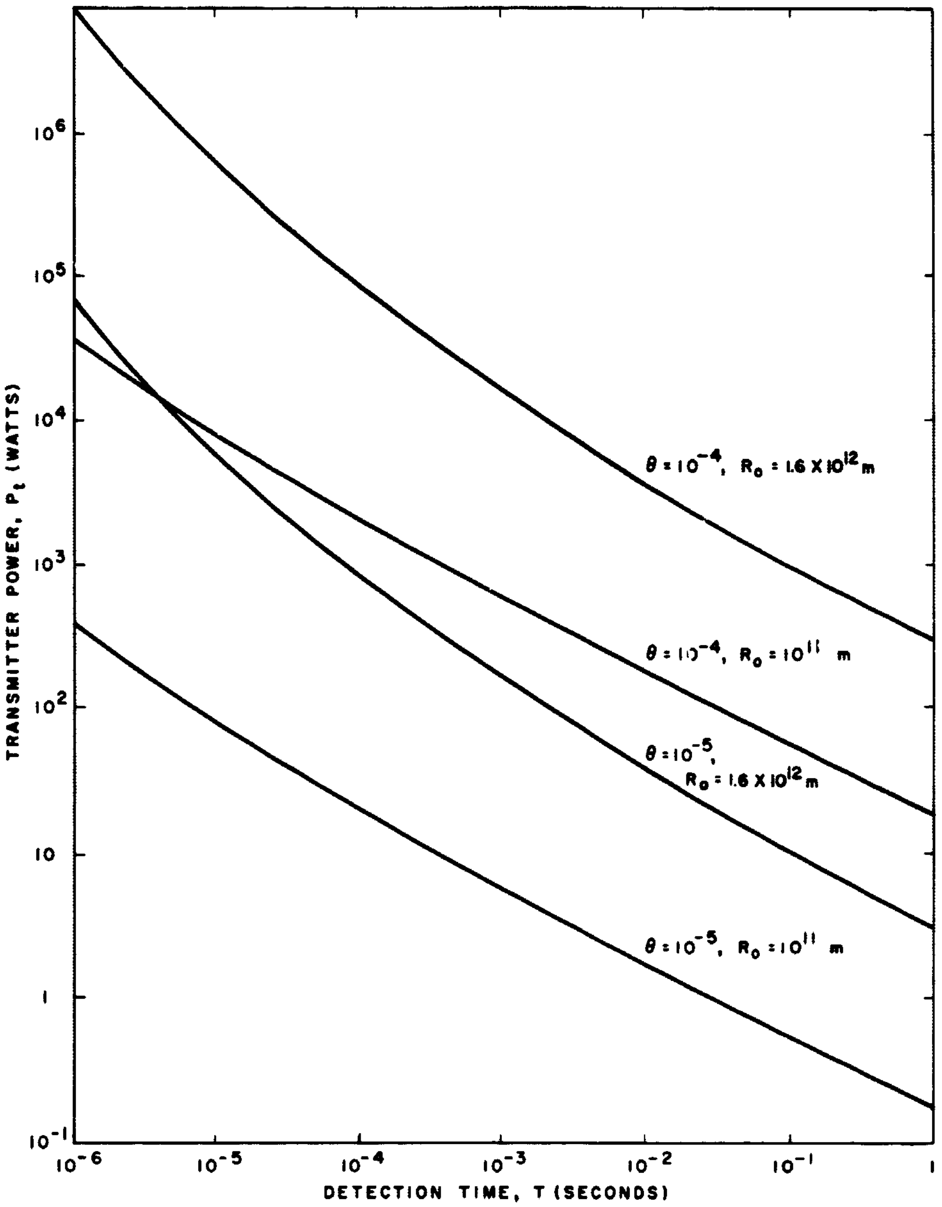


Figure 102. Transmitter power required for beacon acquisition

these quantities and dramatic improvements are not expected.

The above discussion assumes a CW beacon. It is of interest to determine the requirements on a pulsed beacon. In a range of 10^{11} meters, it follows from Figure 102 that a pulsed beacon with a beamwidth of 10^{-4} radian, a peak power less than 0.1 megawatt and a pulse duration of 1 μs would provide sufficient energy for detection above the background. A 0.1 joule pulse is certainly within the capability of pulsed lasers; the difficulty arises in that the laser would have to be pulsed at a very high rate. In the tracking control loop analysis, it is indicated that a rate of 1600 p/s might be required for tracking. Even if the detection filter bandwidth is an order of magnitude narrower perhaps because of a quieter environment in space, a pulse rate in excess of 100 p/s is indicated. This is an average power of about 10 watts and is outside the range of present high power lasers (see Chapter 3).

The situation is somewhat better at a beamwidth of 10^{-5} radian, where the corresponding numbers are somewhat less than a 1kW pulse with a duration of 1 μs. At a rate of 1600 p/s, this corresponds to an average power of about 1 watt and is likely to be achievable. The CW requirement for this narrow beamwidth is about 3 watts, which may be achievable.

It is implicit in the above calculation for the performance of a pulsed beacon system that the receiver is gated about the true arrival. In the acquisition phase this cannot be done, and a larger threshold (and consequently a higher signal level) is required because of the increased number of times at which a false alarm may occur. The increase in threshold is readily estimated. In the absence of signal the detector envelope may be assumed to be Rayleigh distributed. It follows that, if there are M independent times at which a false alarm may occur, then the false alarm probability is given by

$$P_{fa} = 1 - \left(1 - e^{-R_M^2 / 2\sigma^2}\right)^M = e^{-R_1^2 / 2\sigma^2} \quad (21)$$

where R_M is the threshold, σ^2 is the variance of the noise, and R_1 is the threshold that would apply if the receiver were gated. For $P_{fa} << 1$, it follows from Equation (21) that

$$\frac{R_M^2}{R_1^2} = 1 + \frac{\log M}{-\log P_{fa}} \quad (22)$$

For example, for $P_{fa} = 10^{-5}$ and $M = 6(10)^2$ corresponding to a 1 μs pulse with an interpulse time of $6(10)^{-4}$ s) then a 2 dB threshold increase is required. To obtain the same detection probability, the signal power must be increased by approximately the same factor.

It should be noted that the above calculations were all performed for Mars distances. The Earth shine background point intensity, π_ν, decreases a $1/R_0^2$. In Figure 102, the transmitter power as a function of detection time is also plotted for a range of 10 AU(1.6×10^{12} meters). At the shortest detection times the power required varies approximately as R_0^2 since, at these very short times, the system operates in approximately the shot noise limit. At the longer detection times (in the background noise limit), Equation (20) is applicable and the transmitter power required is approximately proportional to R_0. The acquisition of an optical beacon at 10 AU does not appear feasible. At 1600 p/s and $\theta = 10^{-5}$ radian, more than 100 watts average power for a pulsed high-power laser would be required (i.e., about 1600 pulses of nearly 0.1 joule in each second). Even assuming that the pulse rate can be reduced by a factor of 10 leaves the requirements for the laser beyond attainment.

Unlike the usual situation at microwaves, the power requirement on the beacon is more severe than that on the high data rate communications transmitter. This arises largely from the fact that the beacon, which is pointed on an open loop basis, must operate with a considerably broader beam than the narrow-beamwidth spacecraft transmitter.

4.2 Tracking Loop Analysis

The basic limitation in the pointing of a laser beam to high angular precision is the ability of a common optics tracker to track an Earth beacon in the presence of background and shot noise and when disturbed by random accelerations of the telescope system. The acceleration disturbances acting on the transfer lens of the optical system can be reduced by a tight (wide-bandwidth) servo loop. However, the wider the bandwidth of the servo, the more (white) detector noise is allowed to disturb the system. To reduce detector noise a narrow bandwidth is desired. There exists an optimum servo bandwidth to achieve the minimum rms tracking error in the presence of detector noise and disturbing accelerations. This optimum is found in this section as a function of signal-to-noise ratio of the detected beacon signal.

4.2.1 Assumptions

The tracker consists of the main reflecting mirror, a transfer lens positioned by the tracking loop, and an image-dissecting prism which acts as an error detector. The following analysis will be limited to one dimensional motion, because the right-left and up-down systems are uncoupled and dynamically identical. Let θ denote the actual beam direction, θ_i the desired beam direction, and $\vartheta = \theta_i - \theta$ the error in beam direction (Figure 103). The error signal is

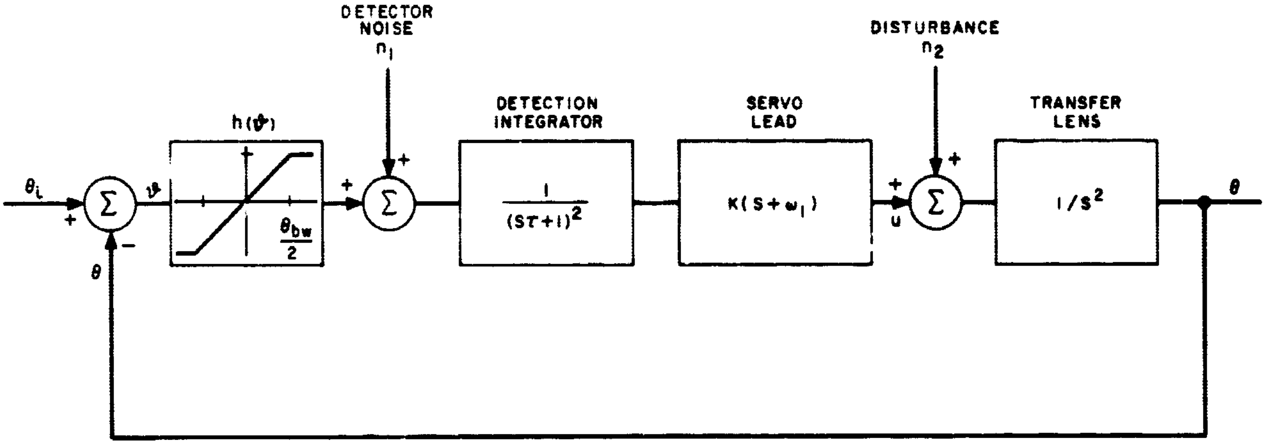


Figure 103. Block diagram of transfer lens servo

processed by focusing the Airy disc of the light signal on the image-dissecting prism to get an error signal that varies as a function of the error $h(\vartheta)$. Associated with the error signal is background and shot noise from the detection process. The detector output is processed by a servo equalizer and filter network with transfer function:

$$G(s) = \frac{K\,(s+\omega_1)^*}{(s\tau+1)^2}$$

The output of the compensating network is denoted as u and drives the transfer lens electromechanical actuator. The dynamics of the transfer lens system is assumed to be a simple inertia, $1/s^2$. This assumes that the electromechanical driver is very fast for the purposes of noise analysis. The block diagram of Figure 103 shows detector noise n_1 and disturbing acceleration n_2 from the motion of the telescope mount, which is itself controlled. Small vibrations induced by moving parts, gas jets, and control motions (limit cycles) will contribute to the acceleration input n_2.

The following further assumptions are made to facilitate an adequate approximate analysis of the steady-state noise response:

1. The detector nonlinearity $h(\vartheta)$ is taken to be linear near $\vartheta = 0$ and reach a completely saturated state corresponding to $\vartheta = \theta_{bw}/2$, where θ_{bw} is the diffraction-limited beamwidth of the main optic. Saturation corresponds to an error signal such that the Airy disc lies entirely to one side of the image dissector. The incremental slope is $(dh/d\vartheta)_{\vartheta=0} = 1$.
2. The detector noise n_1 is assumed to be white (prefiltering) noise. The rms angular noise in a bandwidth W is $\sigma_\theta = \left[(\theta_{bw}) \,/\, (2S_{max})\right] \sigma_n$, where S_{max} is the maximum electrical signal corresponding to full angular signal $\theta_{bw}/2$ and σ_n is the rms post detection electrical noise. The power spectral density of the noise n_1 is then

$$N_\theta = \left(\frac{\theta_{bw}}{2}\right)^2 \left(\frac{N_d}{S^2_{max}}\right)$$

   where N_θ, N_d are the angular and electrical noise power per unit bandwidth [(radian)^2/Hz].
3. The detector integration time τ is small compared to $1/\omega_1$, the lead time constant. The effect of the lag term breaking at $1/\tau$ is neglected in calculating the primary response of the servo. The filter $1/(s\tau+1)^2$ must be used to keep the driver signal u from saturating.
4. The gain K is adjusted for critical damping, $K = 4\omega_1$.
5. The acceleration noise n_2 is assumed to have an exponential autocorrelation function $\phi_2(T) = \sigma_a^2\, e^{-\omega_a |T|}$, where ω_a is the spectral bandwidth and σ_a is the rms acceleration disturbance.
6. The beamwidth θ_{bw} is the same as the transmitted beamwidth of the outgoing laser signal, since the optics are common. The ratio of beamwidth to rms tracking noise is $k = \theta_{bw}/(\overline{\vartheta^2})^{1/2}$ and may be taken to be a constant, for example, k = 10.

4.2.2 Optimum Noise Response

In the linear region ($k > 1$) of $h(\vartheta)$ the response in the s-domain with $\theta_i = 0$ is

$$\vartheta = \left[\frac{G(s)n_1 + n_2}{s^2 + G(s)}\right]$$

$$\approx \frac{4\omega_1\,(s+\omega_1)n_1 + (s\tau+1)^2 n_2}{(s\tau'+1)\,(s\tau''+1)\,(s+2\omega_1)^2}$$

$$u = \frac{G(s)\,\left[s^2 n_1 - n_2\right]}{s^2 + G(s)}$$

$$\approx \frac{4\omega_1\,(s+\omega_1)\,(s^2 n_1 - n_2)}{(s\tau'+1)\,(s\tau''+1)\,(s+2\omega_1)^2}$$

The factoring in the denominators in the forgoing expressions follows (Figure 104) from $\tau\omega_1 \ll 1_a$ and that $\tau'\omega_1$ and $\tau''\omega_1$ remain small compared to one if $K = 4\omega_1$. The lags (approximately τ) provide attenuation in the forward loop to avoid overdriving the transfer lens driver amplifiers and electromechanical actuators.

The spectral density functions corresponding to n_1 and n_2 are:

$$S_1 = N_\theta = \left(\frac{\theta_{bw}}{2}\right)^2 \frac{N_d}{S^2_{max}}$$

$$S_2 = \frac{2\omega_a\sigma_a^2}{(\omega^2+\omega_a^2)}$$

where $s = i\omega$. The variance (zero mean) of ϑ is by Parseval's theorem:

$$(\overline{\vartheta}^2) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega(16N_\theta\omega_1^2)\left\{\left[\frac{\omega_1 + i\omega}{4\omega_1^2-\omega^2+4i\omega\omega_1}\right]\right.$$

$$\left.\left[\frac{\omega_1-i\omega}{4\omega_1^2-\omega^2-4i\omega\omega_1}\right]\right\} + \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\left\{\left[\frac{\sqrt{2\omega_a}\,\sigma_a}{\omega_a-i\omega}\right]\left[\frac{\sqrt{2\omega_a}\,\sigma_a}{\omega_a+i\omega}\right]\right.$$

*s is the Laplace transform complex frequency variable.

$$\left[\frac{1}{(2\omega_1+i\omega)^2}\right]\left[\frac{1}{(2\omega_1-i\omega)^2}\right]\Bigg\}$$

By using the tables on page 372 of Newton, Gould, and Kaiser[52] one obtains the relation:

$$\overline{\vartheta^2} = \frac{5\theta_{bw}^2 N_d \omega_1}{8S_{max}^2} + \frac{\sigma_a^2(4\omega_1+\omega_a)}{16\omega_1^3(2\omega_1+\omega_a)^2}$$

To normalize this expression, set $x = \omega_1/\omega_a$. Using the definition of k gives

$$\frac{\overline{\vartheta^2}}{\sigma_a^2/\omega_a^4} = \frac{(4x+1)}{16x^3(1-x/x_\infty)(2x+1)^2}$$

where $x_\infty = 8S_{max}^2/5K^2N_d\omega_a$. The parameter x_∞ is proportional to the signal-to-noise ratio in the disturbance bandwidth, divided by the square of the ratio of beamwidth to rms tracker noise. The variable $x = \omega_1/\omega_a$ is the ratio of servo bandwidth to disturbance bandwidth.

The response function $\overline{\vartheta^2}/(\sigma_a^2/\omega_a^4)$ goes to infinity at $x = 0$. From $+\infty$ at $x = 0$, the response function decreases to a minimum value denoted $\left[\overline{\vartheta^2}/(\sigma_a^2/\omega_a^4)\right]_{opt}$ at x_{opt}, and then rises to $+\infty$ at $x = x_\infty$. For $x > x_\infty$ there are no positive values of the right-hand side.

In Table 45 the values of $\left[(\overline{\vartheta^2})^{1/2}/(\sigma_a/\omega_a^2)\right]_{opt}$ and x_{opt} are tabulated for various values of x_∞.

Taking the servo system bandwidth as $5\omega_1/4$, the signal-to-noise ratio may be written as $2S_{max}^2/5N_d\omega_1$. Setting k=10 for good overall performance, the value of x_∞ is thus determined. For each value of x_∞, the lead frequency ω_1 is chosen to correspond to the minimum (mean square) angular excursion which is then set equal to $\theta_{bw}/10$ in order to assure tracking the main optical beam. A tabulation of representative values, suitably normalized, is given below.

Table 45

OPTIMUM DESIGN OF THE CONTROL SYSTEM

$x_\infty = \dfrac{8S_{max}^2}{5k^2N_d\omega_a}$	$(\omega_1/\omega_a)_{opt}$	$\left[(\overline{\vartheta^2})^{1/2}/(\sigma_a/\omega_a^2)\right]_{opt}$
0.10	0.075	24.1
0.20	0.151	8.4
0.5	0.383	1.96
1.0	0.776	0.612
2.0	1.57	1.76×10^{-1}
5.0	3.97	3.19×10^{-2}
10.0	7.96	8.33×10^{-3}
20.0	16.0	2.13×10^{-3}
50.0	40.0	0.346×10^{-4}
100.0	80.0	0.86×10^{-4}
200.00	160.0	2.18×10^{-5}
500.00	400.0	3.49×10^{-6}
1000.00	800.0	8.7×10^{-7}

$\left[\theta_{bw}/(\sigma_a/\omega_a^2)\right]_{opt}$	$(\omega_1/\omega_a)_{opt}$	$\left[2S_{max}^2/5N_d\omega_1\right]$
19.6	0.383	32.6
6.12	0.776	32.2
1.76	1.57	31.8
0.319	3.97	31.5
0.0833	7.96	31.4
0.0213	16.0	31.3
0.0035	40.0	31.2

This indicates that a signal-to-noise ratio of 15 dB is required. The bandwidth of course increases with the normalized beamwidth θ_{bw}. For a disturbance (σ_a/ω_a^2) of 50 μ rad (rms) in a bandwidth of f_a = 10 Hz, a servo bandwidth of 160 Hz is required to achieve a θ_{bw} = 1 μ rad.

It has been proposed that a high-energy pulsed laser be used to achieve beacon power at sufficient beamwidth. Such lasers lose single-pulse energy as the pulse repetition rate increases. The servo requires approximately 10 samples/s/Hz of servo bandwidth to operate as an efficient sampled data system with low ripple.[53] For the example given above, the servo bandwidth was 160 Hz. Thus the pulse frequency of the beacon would require 1600 pls. This is quite large for a high-power laser.

4.2.3 Conclusions

The study of the optimum trade-off bandwidth showed that the disturbing accelerations acting on the transfer lens may limit performance in the presence of tracking noise. The bandwidth and signal-to-noise ratio for a typical disturbance environment indicate trouble with the beacon signal power. This study certainly points to the desirability of tests of actual hardware under realistic disturbances and weak signals with background noise.

A more complete study would incorporate more realistic models of the tracker and the transfer lens system. The nonlinearity of the acquisition process could be simulated on an analog computer. Use of the Wiener filter methods of Reference 32 instead of the critical damping criterion could lead to somewhat improved performance. Refinement of the model of disturbances would be desirable.

5. MODULATION

In principle, the use of a coherent optical carrier affords the same flexibility of modulation as a conventional radio carrier. In practice, however, the characteristics of optical

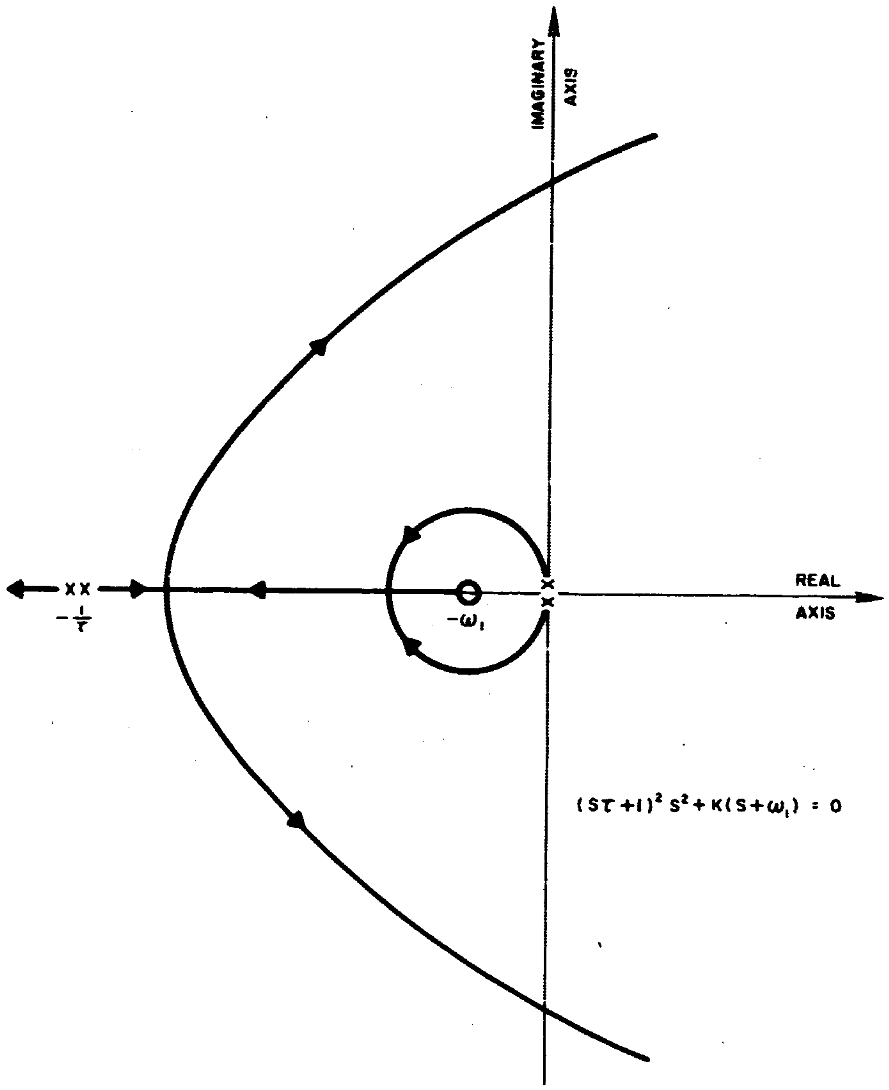


Figure 104. Root locus in S-plane vs. gain K

transmitters, modulators, and detectors, together with propagation effects, tend to make the situation somewhat different at optical than at microwave frequencies. It is the purpose of this section to point out these differences, to indicate their consequences, and to recommend modulation techniques suitable for both heterodyne and direct-detection optical systems.*

In the case of heterodyne detection in an Earth satellite receiver, coherent modulation and detection techniques may be profitably employed. As in the microwave case, maximum efficiency may be obtained with a biorthogonal modulation system in which the signaling waveforms are comprised of phase-reversal sequences. These may be obtained by driving an electro-optic modulator with an appropriate binary waveform, which takes on the values ±V volts, such that a transition of 2V volts results in a phase change of slightly less than 180 degrees in the optical carrier.

The above modulation system requires that the carrier phase remain essentially constant over a pulse period. This requires only that the spectrum of the phase noise, arising either from oscillator fluctuations or atmospherics, be narrow compared with the modulation rates of interest. For megabit communication rates this is easily satisfied, but for kilobit rates and below, this can be a serious problem.

Slow amplitude fluctuations are a more serious problem, because reduced signal power may cause a very significant increase in error probability since then no averaging over the fluctuations takes place in an observation interval. Indeed, if the received signal amplitude is Rayleigh distributed, then the error probability is no longer an exponential decreasing function of E/N_o (the energy per bit normalized by the noise power per unit bandwidth), but decreases as $(E/N_o)^{-1}$. As shown in Section 2, the presence of even weak amplitude fluctuations can cause an appreciable increase in the E/N_o required to achieve a small error probability (see Figure 97).

Thus, in the case of heterodyne reception, modulation considerations are the same as in the microwave case (where phase-shift modulation is the preferred approach) assuming phase fluctuations are slow (valid assumption for most cases of interest) and either amplitude fluctuations are sufficiently weak (questionable assumption for earth reception) or appropriate margin is added.

In the case of direct detection, phase-shift modulation cannot be employed because the detector is insensitive to phase. The choices available here are:

1. Polarization shift modulation
2. Amplitude modulation (on-off)
3. Discrete pulse-position modulation
4. Discrete frequency shift keying.

In binary polarization modulation, information is conveyed by the polarization of the optical signal, i.e., two orthogonal polarizations are employed to convey a 0 and a 1. The choice between orthogonal linear polarization or right- and left-hand circular polarization is one of the convenience rather than of theoretical performance. However, circular polarization has the advantage of being insensitive to misorientation of spacecraft and receiver co-ordinate systems and is also insensitive to Faraday rotation.

In the receiver, the two polarizations are optically separated and separately detected. The decision as to which polarization was transmitted is made on the basis of which polarization channel contains the largest signal in the bit period of interest. The theoretical performance of this orthogonal binary system (see Section 7) is the same as an on-off system which employs the same average power. The on-off system also has the advantage of requiring only a single detector, and no polarization separator is required if background noise is not a problem. On the other hand, and on-off system requires twice the peak power of a CW system to achieve the same average power, However, at a duty factor of 50 percent, a laser transmitter will generally be peak rather than average power limited at the large pulse rates required, so that the on-off system has a 3 dB penalty relative to the polarization system. Also, in the polarization system, a comaprison is made between two channels, rather than comparison with a threshold as in the on-off system. The latter technique is much more sensitive to amplitude fluctuations caused by atmospheric attenuation. In addition, for tracking and synchronization purposes, it would be necessary to guard against a long string of "offs." For all these reasons, polarization modulation is clearly superior to amplitude modulation.

Binary frequency-shift keying (FSK) is likewise an orthogonal system and has the same theoretical performance as binary polarization modulation. Moreover, M-ary frequency shift may be employed to achieve an M-ary orthogonal system with theoretical performance approaching that of biorthogonal (Chapter 1, Section 6) systems. There are, however, two difficulties with this approach, one practical and the other theoretical. From the pratical standpoint in a direct-detection FSK system it is necessary to employ frequency shifts greater than the resolution capability of optical filters. Such frequency shifts† typically exceed the line-width capability and tuning range of a single laser, and hence would necessitate multiple lasers.

A more fundamental problem arises in the detector. In an M-ary FSK receiver, the incoming signal is divided into M channels, each of which contains a bandpass filter centered at the appropriate frequency. In a conventional microwave system, this power division results in a decrease

*As discussed in Chapter 1, Section 6, attention will be restricted to digital modulation systems. In the case of optical systems, this affords the additional advantage of relaxing linearity requirements on the modulator.

†Note that a 1 Å filter width at λ = 0.5 micron corresponds to a bandwidth of $1.2\ (10)^{11}$ Hz.

in the signal level accompanied by a corresponding decrease in noise, so that the signal-to-noise ratio is unchanged. However, in an optical system which operates at the quantum limit, a reduction in signal level due to power division results in a relative increase in shot noise. Thus a binary optical FSK system which divides the incoming signal equally between the two channels suffers a 3 dB penalty relative to a polarization system in which virtually all of the signal may be recovered in the appropriate polarization channel.

Direct-detection optical systems cannot afford the luxury of power division prior to the detector. Although it is feasible in principle to build optical FSK receivers based on frequency-sensitive beam-deflection techniques, it is not expected that the losses associated with such techniques would be sufficiently low to justify their use. Thus, both for reasons of implementation and in terms of expected performance, FSK systems are less attractive than binary polarization modulation.

The use of discrete pulse-position modulation permits the achievement of M-ary orthogonal systems without the power-division problem associated with FSK systems. Also, the modulation problem–viz., controlling the time at which the laser is pulsed–is simpler. For PPM systems to have marked superiority to polarization modulation, comparable average power must be radiated and duty factor must be low, implying large peak power. There is still the requirement for high average pulse-repetition frequency (high data rate requirement) and variable interpulse times (because of the modulation). This may not be practical with contemplated laser systems and will incur considerable complexity at best. (Numerical estimates of potential benefit of PPM systems are given in Appendix 9.)

By the process of elimination, it appears that binary polarization modulation is a likely choice for direct-detection optical systems.

6. HETERODYNE DETECTION

The use of optical heterodyne detection systems may be advantageous in both satellite-borne and ground receivers. For a satellite receiver, where it might be anticipated that in the absence of an atmospheric path there is negligible disturbance in the signal beam, one can consider use of coherent modulation as in conventional communication at longer wavelengths. This will be possible as long as the laser frequency itself is sufficiently stable. When an atmospheric path is involved, the maintenance of relative temporal phase between signal and local fields for appreciable periods of time is uncertain.* Nevertheless, the heterodyne principle may still be profitably employed for incoherent modulation. The benefits to be derived in so doing are the conversion gain, which minimizes the effect of detector noises, and the improved (relative to direct detection) discrimination against background noise. The latter discrimination derives from the well-known fact that for the mixing of waves spatial phase coherence must be maintained, hence only the background radiation which is in the same transverse mode(s) as the local oscillator radiation will mix with it.

6.1 Heterodyne Signal-to-Noise Ratio

The combined local, signal, and background fields excite a photocurrent which undergoes an essentially noiseless multiplication M to form the photomultiplier output. After filtering, the current in the IF band may be written as

$$i_{IF} = M(i_{so}+i_{bo}+i_{sb}+i_{bb}+i_{ns}+i_{nb}+i_{no}+i_{nd}+i_{th/M}) \qquad (23)$$

These terms are defined as:

i_{so} = current due to mixing of signal and local oscillator (LO)

i_{bo} = current due to mixing of background and LO

i_{sb} = current due to mixing of signal and background

i_{bb} = current due to self-mixing of background field

i_{ns} = shot noise current generated by signal

i_{nb} = shot noise current generated by background

i_{no} = shot noise current generated by LO

i_{nd} = shot noise component of dark current

i_{th} = effective thermal noise current in IF band

Here an ideal local oscillator is assumed which introduces no noise other than a contribution to the shot noise: i.e., there are no LO sidebands. Biernson and Lucy[54] have carefully enumerated the various contributions to the noise (except for the background self-mixing) and the reader is referred to their work for a more complete discussion.

Forming the signal-to-noise ratio, again after Biernson and Lucy

$$\mathrm{SNR} = \frac{\overline{i_{so}^2}}{\overline{i_{bo}^2} + \overline{i_{sb}^2} + \overline{i_{bb}^2} + \overline{i_{ns}^2} + \overline{i_{nb}^2} + \overline{i_{no}^2} + \overline{i_{nd}^2} + \overline{i_{th/M^2}^2}} \qquad (24)$$

Expressing the mean-squared currents in Equation (24) in terms of the collected optical powers and other system parameters leads to

$$\mathrm{SNR} = \frac{2C^2P_sP_o}{2C^2(P_{bh}P_o + P_{bh}P_s + P_{bh}P_b) + \frac{4kT_aB}{R_aM^2} + 2eBC(P_o + P_s + P_b + P_d)} \qquad (25)$$

*If the relative phase is changing slowly with respect to the bit rate, measurement of this relative phase is possible and in principle, would permit use of temporally coherent modulation.

The quantities appearing in Equation (25) are defined as follows:

$$c = \frac{\eta\tau_o e}{h\nu}$$

where h = Planck's constant ($6.63\text{x}10^{-34}$ Joule-sec)

ν = Frequency of radiation

η = quantum efficiency of photosurface

τ_o = Transmissivity of predetection optical filter

e = Electronic charge ($1.6\text{x}10^{-19}$ coulombs)

k = Boltzmann's constant ($1.38\text{x}10^{-23}$ Joule/°K)

T_a = Noise temperature of IF amplifier (°K)

R_a = Input resistance of IF amplifier

B = Bandwidth of IF filter

P_s = Signal power collected by receiving aperture

P_o = Local oscillator power collected by receiving aperture

P_b = Background power generating shot noise

P_{bh} = Background power translated into IF band by LO

P_d = Equivalent optical input power giving rise to dark current.

Equation (25) may be simplified to read

$$\text{SNR} = \frac{P_s}{P_{bh}\left(1 + \frac{P_s}{P_o} + \frac{P_b}{P_o}\right) + \frac{1}{P_o}\,\frac{2kT_aB}{M^2R_a}\left(\frac{h\nu}{\eta\tau_o e}\right)^2 + \frac{h\nu B}{\eta\tau_o}\left(1 + \frac{P_n}{P_o}\right)} \quad (26)$$

where

$$P_n = P_s + P_d + P_b$$

If P_o can be made sufficiently large, and if P_{bh} is small, Equation (26) becomes

$$\text{SNR} = \frac{\eta\tau_o P_s}{h\nu B} \quad (27)$$

which defines the ideal heterodyne response, limited only by the local oscillator shot noise.

The analysis leading to Equations (26) and (27) must be altered slightly before it can be used to describe photodetection in the 10μ wavelength region (see Chapter 3, Section 7). Modification is necessary because photomultipliers are relatively ineffective in this range; photoconductive detectors are found to be most useful at or near 10μ wavelengths. One need note that, for the photoconductive detector, $M = 1$ and if a dark current is defined it must be an equivalent quantity. That is, for these devices, it is difficult to separate the various contributions to the residual noise. However, an equivalent dark current may be defined whose resultant shot noise* is equal to the combined thermal noise, background shot noise, and shot noise induced by the bias current. The quivalent dark current and its corresponding optical input P_d can be calculated for any given photoconducting detector from its specified "noise-equivalent-power" (NEP). Consequently, one may use Equation (26) at 10μ if one drops the thermal noise† term and also the background term from P_n, and recognizes that these will be accounted for by the equivalent P_d component of P_n.

6.2 Operating at the Shot-Noise Limit

When the first two terms of the denominator of Equation (26) are dominated by the last term, the limiting response described by Equation (27) may be approached. However, before the limiting response can be achieved, it is also necessary that

$$\frac{P_n}{P_o} \ll 1 \quad (28)$$

which follows if this inequality holds for each of the terms comprising P_n; i.e., if

$$\frac{P_s}{P_o} \ll 1 \text{ or } P_o \gg P_s = \pi_s A_R \quad (29)$$

$$\frac{P_b}{P_o} \ll 1 \text{ or } P_o \gg P_b = N_\nu \Omega_R A_R W \quad (30)$$

$$\frac{P_d}{P_o} \ll 1 \text{ or } P_o \gg P_d = \frac{h\nu}{\eta\tau_o e} I_d \quad (31)$$

In these expressions, π_s is the signal power density at the photosurface, A_R is the area and Ω_R the field of view of the receiving aperture, N_ν is the background irradiance discussed in Section 1.1, W is the bandwidth of the predetection optical filter, and I_d is the average dark current emitted by the photosurface. The quantities appearing on the right in the above expressions have been computed for some typical system parameters.

Received signal power levels have been estimated, assuming a transmitter-receiver separation of 10^{11}m, and the estimates are shown in Table 46.

*In semiconductor nomenclature, shot noise is the generation-recombination (g-r) noise.

†This assumes that the specified value of NEP is for a measurement corresponding to conditions; i.e., detector temperature and background level, which are of interest in the application under consideration.

The information in Table 40 (Section 1.1) was used in computing the background flux due to atmospherically scattered sunlight and direct thermal radiation, quantities which are necessary for the computation of P_b shown in Table 47.

Table 48 contains values of dark current and the corresponding equivalent optical input power for some commercially available photodetectors which can be used at the wavelengths of interest.

From the data in the preceding tables, the most intense source of shot noise in the 0.5 to 1.0μ range is of the order of 10^{-8}w. At 10μ the equivalent dark-current input power is considerably greater being in excess of 2 milliwatts for the detector considered. As a result, even a modest sourçe of LO power is sufficient to insure that the shot noise will be dominated by the LO. Submicrowatt levels will suffice in the visible, while milliwatts, or even some tens of milliwatts, may be required at 10μ. Having observed that a local oscillator dominated shot noise condition can be reached under reasonable circumstances, the effective local oscillator shot noise, hr $B/\eta\tau_o$, must be compared with the first two terms in the denominator of Equation (26). The comparison appears in Table 49.

Table 46

RECEIVED SIGNAL POWER

λ	Ω^*_T (steradian)	Transmitted Power (watts)	π_s	$\pi_s A_R$ ($A_R = 1m^2$) (watts)
0.5μ	10^{-12}	1 (Argon ion laser)	10^{-10} w/m²(cw)	10^{-10}
1.0μ	$4X10^{-12}$	1υ0 (Nd:YAG laser)	$2.5X10^{-9}$ w/m²(cw)	$2.5X10^{-9}$
		1k (pulsed at 5 5 kHz)	$2.5X10^{-8}$ w/m²	$2.5X.10^{-8}$
10.0μ	$4X10^{-10}$	10 (CO_2 laser)	$2.5X10^{-12}$ w/m²(cw)	$2.5X10^{-12}$

*Transmitter beam divergence taken as 4X diffraction limit: $\Omega_T = 4X(\lambda/D)^2$.

Table 47

EFFECTIVE BACKGROUND NOISE POWER

λ	N_ν	Ω^*_R (steradian)	A_R (m²)	W (Å)	$P_b = N_\nu \Omega_R A_R W$ (watts)
0.5μ	$5.7X10^{-14}$ w/m²–ster–Hz	10^{-12}	1	0.1	$6.9X10^{-16}$
			1	1.0	$6.9X10^{-15}$
	Atmosphere Scattered Sunlight	10^{-8} (atmosphere limited)	1	0.1	$6.9X10^{-12}$
			1	1.0	$6.9X10^{-11}$
1μ	$8.0X10^{-14}$ w/m²–ster–Hz	$4X10^{-12}$	1	0.1	$9.6X10^{-16}$
			1	1.0	$9.6X10^{-15}$
	Atmosphere Scattered Sunlight	$4X10^{-8}$ (atmosphere limited)	1	0.1	$9.6X10^{-12}$
			1	1.0	$9.6X10^{-11}$
10μ	$4.4X10^{-10}$ w/m²–ster–Hz	$4X10^{-10}$	1	10	$5.3X10^{-12}$
			1	750	$3.9X10^{-10}$
	Direct Infrarad Radiation From Atmosphere at 300°K	$4X10^{-8}$ (atmosphere limited)	1	10	$5.3X10^{-10}$
			1	750	$3.9X10^{-8}$

*Receiver field of view taken as 4X diffraction limit, $4X(\lambda/D)^2$, except when limited by atmospheric turbulence.

Table 48

PHOTODETECTOR DARK CURRENT

λ	Device	Surface	η	M	I_d	$P_d = \frac{h\nu}{\eta\tau_o e} I_d$ (watts)
0.5μ	7265 (PMT)	S-20	0.18	2×10^{7}	1.2×10^{-13}a (at 25°C)	1.5×10^{-12}w
1.0μ	7102 (PMT)	S-1	0.0036	1.5×10^{5}	4.5×10^{-11}a (at 25°C)	1.6×10^{-8}w
10.0μ	Ge:Cu (photoconductor)	–	0.5	1.0	9.4×10^{-3}a (at 4°K)	2.3×10^{-3}w

Table 49

A COMPARISON OF THE NOISE TERMS IN THE DENOMINATOR OF EQUATION (26)

λ	Ω_R	P_{bh}*	$\frac{h\nu B}{\eta\tau_o}$	$\frac{2kT_aB}{P_oM^2R_a}\left(\frac{h\nu}{\eta\tau_o e}\right)^2$
$0.5\,\mu$	10^{-12} ster.	1.1×10^{-17}w	2.2×10^{-10}w	$\frac{\sim1.6\times10^{-24}}{P_oR_a}$ w
1.0μ	4×10^{-12} ster.	6.4×10^{-17}w	5.5×10^{-9}w	$\frac{1.8\times10^{-17}}{P_oR_a}$ w
10μ	4×10^{-10} ster.	3.5×10^{-13}w	4×10^{-12}w	Included in equivalent dark current

*P_{bh} is that part of the background power contained in one transverse mode as seen by the receiver which lies in the twc temporal sidebands of total width 2B. Hence, assuming the spectrum is flat over a narrow region,

$$P_{bh} = N_\nu(2B)\Omega_R A_R$$

The entries in Table 49 assume an IF bandwidth B of 100 MHz (10^8) and that $\tau_o = 1$. Since it is unlikely that a value of R_a less than a few tens of hundreds of ohms would be used, submicrowatt LO power levels are adequate to reach† the optimum performance indicated by Equation (27) at both 0.5 and 1μ. At 10μ, heterodyned background radiation is considerably more significant but still well below the dominant shot noise.

6.3 Notions of Importance in Heterodyne Detection

The formal development of the expression for the signal-to-noise ratio given in Section 6.1 is strictly valid only for quasimonochromatic fields which closely approximate plane waves at the collecting aperture (assuming one uses a classical description of the photoelectric interaction). Furthermore, the local and signal fields must be perfectly aligned: the wave vectors must be colinear.

It is important to recognize that the mixing terms (e.g., i_{so}) entering Equation (29) represent interactions between the various components of the aperture field. The mean-square mixing current i_{so}^2, for example, is proportional to P_sP_o as assumed in Equation (25) only when the signal and local fields are uniform in intensity and display perfect

†In computing the values shown in the last column of Table 49, the rated PMT multiplications (see Table 48) were used. However, operation at the required local oscillator power levels will necessitate a reduction in gain if the manufacturer's maximum anode current ratings are not to be exceeded. It is not difficult to show that the required reductions can be affected without upsetting the desired shot noise limited condition (utilizing the full PMT gain is not necessary for shot noise limited operation).

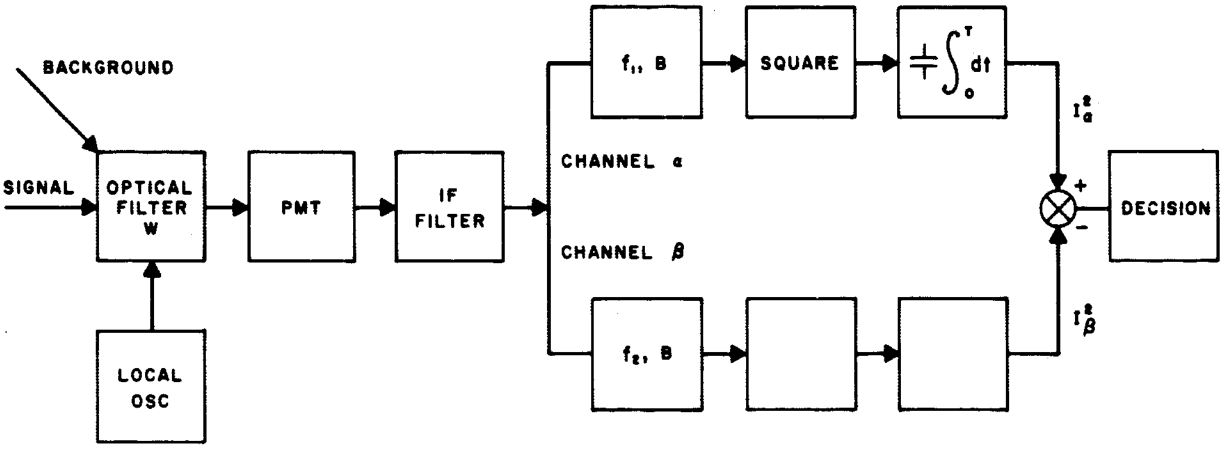


Figure 105. Schematic representation of an optical heterodyne receiver for use in an FSK communication system

spatial coherence over the receiver aperture. In general, a mean-square mixing curren\ is given by an expression such as the following for $\overline{i^2_{so}}$:

$$\overline{i^2_{so}} = 2\left(\frac{\eta\tau_o e}{h\nu Z}\right)^2 \int_A \int_A \overline{x_s x_s'}\ \overline{x_o x_o'}\ d^2\vec{r}\, d^2\vec{r}' \qquad (32)$$

Here Z is the intrinsic impedance of the medium surrounding the photosurfa:e $Z = \sqrt{\mu/\epsilon}$ and $x_s = x_s(\vec{r})$ and $x_s' = x_s(\vec{r}')$ are the electric field amplitudes of the signal (see Appendix 6) at $\vec{r}$ and $\vec{r}$.' Similar notation is used for the local oscillator field amplitudes and signal and LO fields are assumed statistically independent. The double integral is evaluated over the receiver aperture, or the plane of the photosurface.[39]

The integrand in Equation (32) is a product of correlation functions and each factor is significant only when $\vec{r}$ and $\vec{r}'$ are separated by less than the coherence distance of the appropriate field. This means that the evaluation of Equation (32) will lead to

$$\overline{i^2_{so}} = 2\left(\frac{\eta\tau_o e}{h\nu}\right)^2 \pi_s \pi_o A_R \min(a_s, a_o) F \qquad (33)$$

where π_s and π_o are representative power densities and a_s and a_o are the coherence areas of the signal and local fields respectively. F is a factor dependent on the beam geometries and power distributions (the case of Gaussian beams is treated in Appendix 6). Since the coherence areas may be very small, Equation (33) indicates that the mixing current will be seriously limited if either field displays poor spatial coherence properties.

Equation (32) and its analogues are further based on the assumption that the interacting fields propagate colinearly. When this is not the case, the integrand in Equation (32) must be modified by the inclusion of a factor which accounts for relative carrier dephasing across the aperture. The unavoidable presence of this term strongly affects the angular field of view of the heterodyne receiver. When ideal plane waves are being heterodyned, the field of view Ω_R is known to vary as[56]

$$\Omega_R = \frac{\lambda^2}{A_R} \qquad (34)$$

However, when the fields display imperfect spatial coherence, the relation given by Equation (34) must be replaced by an effective field of view

$$\Omega = \frac{\lambda^2}{\min(a_s, a_o)} \qquad (35)$$

which may represent a sizable increase over the former expression. Observe that Equation (35) represents a reduction in local and signal beam alignment tolerance, but at the same time implies a weaker interaction (smaller $\overline{i_{so}}$) between these fields.

6.4 Heterodyne Communication System

Consider a simple (temporally incoherent) communication system employing a heterodyne receiver, a schematic of which is found in Figure 105.* While the modulation could be any binary orthogonal modulation (such as polarization shift modulation) consider a frequency shift system: The incoming signal consists of one of two orthogonal frequencies (or frequency bands). The receiver is assumed to be in perfect synchronization with the transmitter. After the incoming radiation passes through an optical filter W and is detected, the resultant currents, pass through parallel IF filters with pass bands B at the nominal frequencies f_1, f_2 (corresponding to the modulation). The filtered IF currents are then squared and time averaged for an interval T. The two channel outputs are then compared; the decision on the transmitted beam frequency is based on the sign of the difference of the outputs.

One can compute the mean $\overline{I^2}$ and variance $\sigma^2_{I^2}$ of the random variable

$$I^2 = \frac{1}{T}\int_0^T i^2(t)dt$$

Assume that $\sigma^2_{I^2} \approx \frac{1}{2BT}\sigma^2_{i^2}$ and that i_{no} is a zero mean Gaussian random variable which is independent of i_{so}. From this, the mean and variance of the test statistic ΔI^2 ($\equiv I^2\alpha - I^2\beta$) can be found to be

$$\overline{\Delta I^2} = \overline{i^2_{so}}$$

and

$$\sigma^2_{\Delta I^2} = \frac{1}{2BT}\left(\sigma^2_{i^2_{so}} + 4\,\overline{i^2_{so}}\ \overline{i^2_{no}} + 4\,\overline{i^2_{no}}^2\right)$$

If one assumes that ΔI^2 is well approximated as a Gaussian random variable, the integration time (or bit period) can be found from the equation

$$\frac{[\overline{\Delta I^2}]^2}{\sigma^2_{\Delta I^2}} = K$$

Here, $\overline{\Delta I^2}/\sigma_{\Delta I^2}$ may be usefully regarded as the communication signal-to-noise ratio. The quantity K is the solution of

$$P_\epsilon = \Phi(-\sqrt{K})$$

P_ϵ being the desired error probability and Φ the tabulated function defined by

$$\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx$$

*This block diagram will serve to guide the analysis but certain features will be subsequently modified.

For an error probability of 10^{-3}, K, which varies slowly with P_ϵ, is about 10 so that the information rate H = 1/T is about 10 so that the information rate H = 1/T is

$$H \approx \frac{2B}{10}\left(\frac{\overline{i^2_{so}}^{\,2}}{\sigma^2_{i^2_{so}} + 4\,\overline{i^2_{so}}\;\overline{i^2_{no}} + 4\,\overline{i^2_{no}}^{\,2}}\right)$$

or

$$H \approx 0.05B \; \frac{\overline{i^2_{so}} / \overline{i^2_{no}}}{1 + \overline{i^2_{no}} / \overline{i^2_{so}} + \frac{1}{4}\,\frac{\sigma^2_{i^2_{so}}}{\overline{i^2_{so}}\;\overline{i^2_{no}}}} \qquad (36)$$

If the relative phase between signal and local oscillator is virtually unchanged over a bit period and if amplitude fluctuations imposed by the atmosphere are negligible (Section 2.1.5), the last term in the denominator goes to zero. In that case Equation (36) may be expressed in terms of the optical powers

$$H \approx 0.05B \, \frac{\eta\tau_0 P_s / h\nu B}{1 + h\nu B / \eta\tau_0 P_s} \qquad (37)$$

From this equation, it can be seen that the term in the denominator which causes a deviation from the "shot noise limit," $H \propto \eta\tau_0 P_s/h\nu$, becomes unimportant when $B \lesssim \eta\tau_0 P_s/h\nu$; hence, it is desirable to make B small.

Implementing the schematic diagram in detail implies transmitting pulses of two different carrier frequencies, entailing rapid tuning of the transmitter (or switching between two transmitters). Alternatively, the frequency shifting can be introduced by modulating the phase of the carrier in accordance with the information to be transmitted. This can be done with a small-deviation FM. In this case the carrier is transmitted along with information-carrying side bands. Then at the receiver a wideband IF filter can be followed by a carrier tracking filter using a voltage-controlled oscillator (actually tracking the heterodyned carrier), and the measured frequency can be mixed with the information-carrying bands. The spectral density of the transmitted carrier may be made large to achieve accurate tracking while the width can be kept narrow so that little power is wasted in transmitting this reference. The filters at f_1, f_2 can assume a very narrow width, as long as the carrier frequency does not change appreciably during an information interval. Therefore, the bandwidth in Equation (37) can be made as small as the information bandwidth H without losing information. Hence, Equation (37) may be written

$$H \approx 0.05\,H \, \frac{\eta\tau_0 P_s / h\nu H}{1 + h\nu H / \eta\tau_0 P_s}$$

This implies that $\eta\tau_0\,P_s/h\nu H$ is approximately 20, so that one achieves the shot noise limit*

$$H \approx 0.05 \eta\tau_0 P_s / h\nu \qquad (38)$$

One further point: in this scheme if the doppler shifting of the carrier (due to motion of the transmitter) is slow with respect to the information rate, the IF filter can be made sufficiently broad to accommodate the frequency excursions due to doppler shifting and no tuning of the local oscillator will be necessary.

In Figures 106 and 107 are plotted estimates, based on the shot-noise limit, of transmitter power required as a function of information rate for various circumstances. In the visible region (Figure 106), 0.5μ, 1.0μ it has been assumed that $\eta\tau_0 = 10^{-2}$ and that coherence length is 10 cm for a ground receiver and exceeds the assumed 1 meter receiver aperture for a satellite receiver. At the infrared wavelength (Figure 107) 10μ, $\eta\tau_0$ has been taken as 10^{-1} while the ground level coherence length assumed is 1 meter. In all cases, the length of the communication link has been taken as 10^8 miles, and the transmitter is assumed to be 1 meter in diameter and diffraction-limited. No atmospheric attenuation was included in the calculation.

It can be seen from an examination of Figure 106 that, even at the shot-noise limit, transmission at high data rates from deep space to a ground receiver, at visible frequencies using heterodyne detection, requires transmitter powers that are not practical for the foreseeable future. This is principally due to the limitation of useful receiving aperture imposed by small coherence area. Almost certainly, performance will be improved somewhat by detector development, quite possibly by an order of magnitude. Nevertheless, the use of heterodyne detection in the visible, for information rates exceeding 1 megabit per second and distances exceeding 1 AU seems to hold little promise for a ground receiver.

At 10 microns, the situation is somewhat better due to the (anticipated) larger coherence area and the use of detectors which are quite efficient. However, it should be remembered (Section 2) that the estimates of coherence area (and other associated data) are as yet without experimental verification. Estimates of the performance and reliability of a 10-micron heterodyne communication system from deep space to a ground receiver are crucially dependent on knowledge, such as amplitude fading and spatial phase stability, which is yet to be obtained.

*Since the product of bandwidth and integration time is now unity, one can no longer appeal to the argument that the output of the integrator is approximately Gaussian. This results in a small change in the constant, which has been ignored in Equation (38).

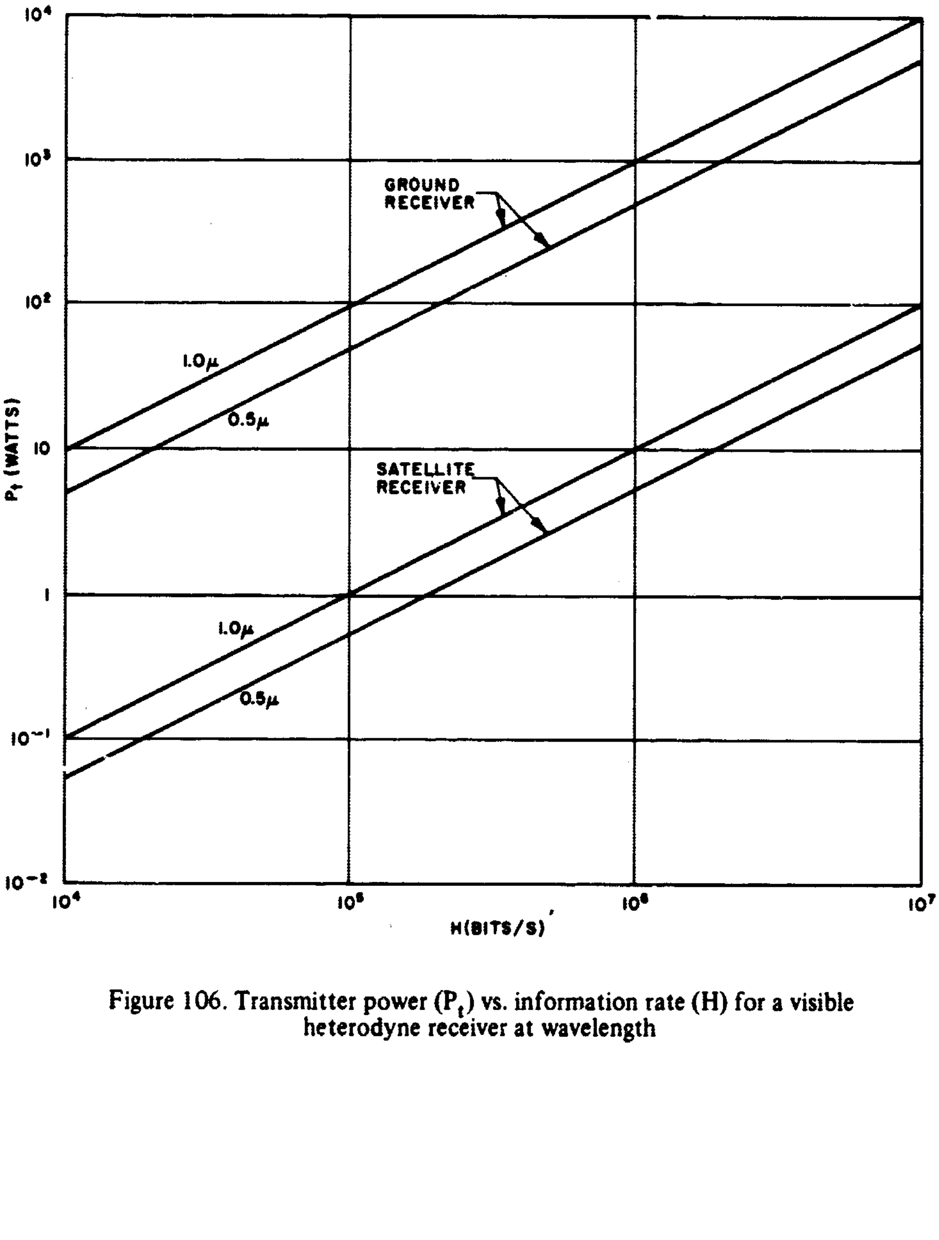


Figure 106. Transmitter power (P_t) vs. information rate (H) for a visible heterodyne receiver at wavelength

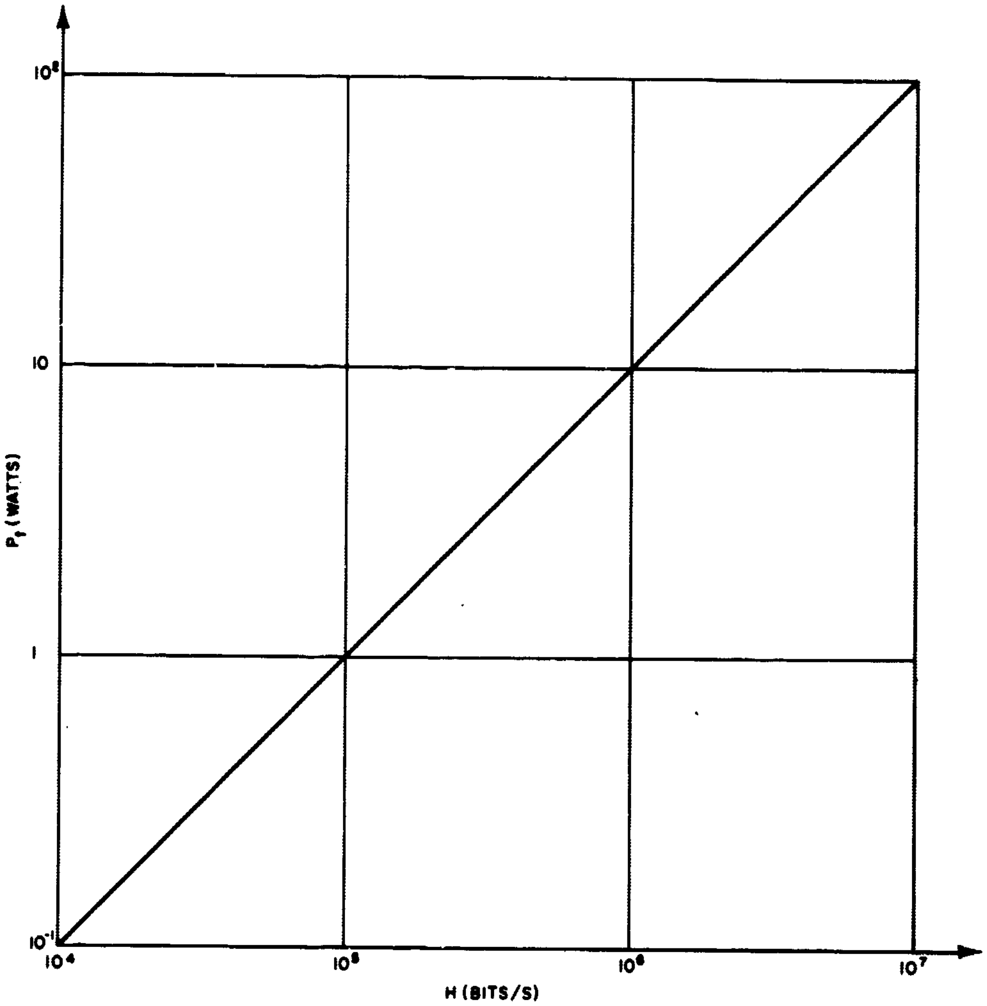


Figure 107. Transmitter power (P_t) vs. information rate (H) fo: a 10μ heterodyne receiver

6.5 Homodyne Detection

The homodyne receiver resembles the heterodyne in the sense that it too employs a locally generated field which, when mixed with an incoming signal field, yields a conversion gain and others of the advantages associated with the heterodyne. In contrast to the heterodyne, however, the nominal local and signal frequencies are the same; thus the post-detection filtering is performed at baseband.

The theoretical advantages of the homodyne have been pointed out by Oliver[58] and others.[59] The principal advantage over the heterodyne is an increase in conversion gain for given signal and local oscillator powers. This happens ideally when the signal and local fields are in phase so that their product (on conversion) has average power which is just the product of their amplitudes. In the heterodyne, by contrast, the mixing term is periodic, and hence has average value half that of the product of amplitudes. The potential 3 dB advantage which in this way accrues to the homodyne system implies the existence of a phase-locking scheme to accomplish this for two oscillators. The feasibility of such a system has been demonstrated experimentally.[60]

Interest in the homodyne system also stems from its attractiveness to the experimentalist. The heterodyne system requires precise control of the frequency offset of the local oscillator with respect to the signal field, and the homodyne provides a convenient way of avoiding this stability problem in the laboratory. One may simply divide the output of a single laser and use one portion as the signal beam and the other as a local oscillator.[61] Prior to the development of useful laser afc systems, this technique, with the addition of a frequency translator in the LO leg, was in fact utilized to perform heterodyne experiments.[27]

If the signal is transmitted through the atmosphere and there is disturbance of the temporal phase by atmospheric effects, the relative phase between the signal and local oscillator is likewise disturbed. However, as is pointed out in Section 2.1.5, the atmospheric effects will generally be governed by a time scale which is long compared to the inverse of the information bandwidth. That is, the atmosphere will remain "frozen" for many bit periods. For example, if the information rate is 1 MHz and the spectrum of atmospheric processes does not extend beyond 1 KHz, then it should be possible to measure the phase of the incoming signal (with an attendant negligible loss in the rate of desired information) and adjust the phase of the local oscillator to achieve the maximum signal.

In the following section the performance of an idealized incoherent digital binary communication system will be considered. Indeed, if the assumption made herein about the slow rate of change of transmission conditions is valid, it would be possible to use coherent modulation. (In fact, the same is true for a heterodyne receiver.) Nevertheless, as discussed previously, digital modulation is easier to implement and also offers flexibility of coding techniques. Furthermore, the idealized performance of a digital system approximates that of coherent systems (in terms of signal-to-noise ratio) and adequately illuminates the physical limitations on either category of system.

6.5.1 Homodyne System Performance

Consider the performance of a binary polarization shift modulation system. (Actually, the homodyne receiver might be implemented with a number of comparable binary orthogonal modulation schemes with the same performance). A block diagram is shown in Figure 108. A pulse of duration T is transmitted by the signal oscillator in one of two orthogonal polarizations. At the receiver, the incoming signal and background radiation are optically filtered (to reduce the background self-beating contributions), then passed through a polarization separator to be separated into the two polarizations. In each of the channels of the receiver, a single component of signal plus noise is mixed with the local oscillator of appropriate polarization. Each detector output current passes through a baseband filter of width B. Then the current is squared and this is time-averaged for an interval T, which is assumed to be perfectly synchronized with the transmitter. The decision as to which polarization was transmitted is made on the basis of which channel has the larger output.

While a more detailed analysis of the performance is presented in Appendix 7, the result will be derived heuristically here. Assuming (as in the analysis of heterodyne performance) that the local oscillator has sufficient power, that the background degeneracy is sufficiently small, and that B is sufficiently small, the remaining baseband currents (assuming the transmitted pulse is in the polarization denoted by a) are

$$i_a = i_o + i_{so} + i_{no}$$

The laser spectra are so narrow and the atmosphere is assumed sufficiently quiet over the pulse period T that i_o and i_{so} can be considered as deterministic. In channel β, where the signal is absent,

$$i_\beta = i_o + i_{no}$$

The mean* difference in the output

$$\Delta I^2 \equiv \frac{1}{T}\int_O^T i_a^2 dt - \frac{1}{T}\int_O^T i_\beta^2 dt$$

*Note that this averaging is only over the ensemble of shot-noise realizations.

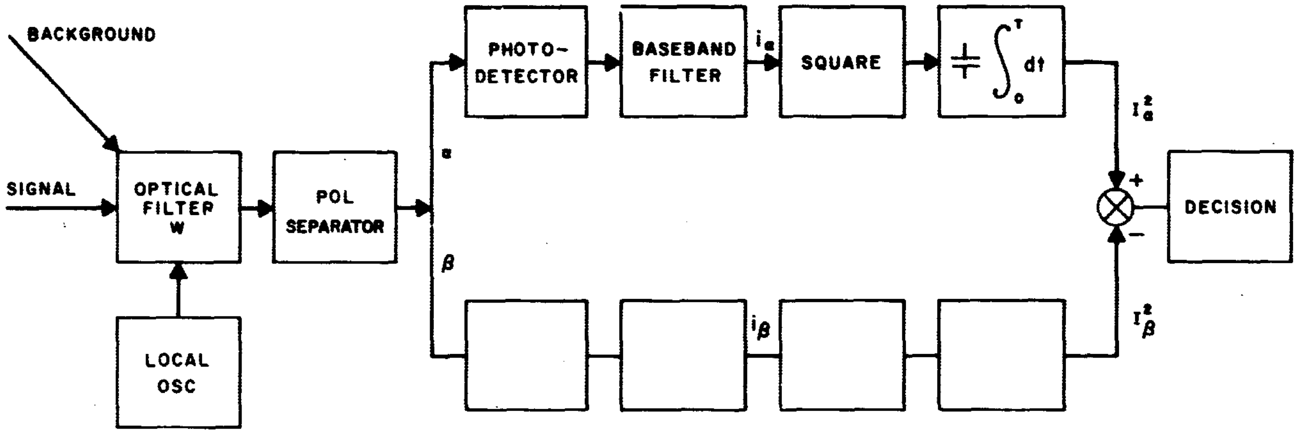


Figure 108. Schematic representation of an optical homodyne receiver for use in a polarization modulation communication system

is then

$$\overline{\Delta I^2} = \overline{i_\alpha^2} - \overline{i_\beta^2} = i_{so}^2 + 2i_{so}i_o$$

For a strong local oscillator

$$\overline{\Delta I^2} \approx 2i_{so}i_o$$

To compute the variance of ΔI^2, note that

$$i_\alpha^2 \approx \left[i_o^2 + i_{so}^2 + 2i_o i_{so} \right] + 2i_{no}(i_o + i_{so})$$

The term in brackets is deterministic in this model; the second term is dominated by $2i_o i_{no}$. Thus

$$\sigma^2_{i_\alpha^2} \approx 4i_o^2 \sigma^2_{i_{no}}$$

The same is true of $\dot{\sigma}^2{}_{i_\beta^2}$. Hence

$$\sigma^2_{\Delta I^2} = \sigma^2_{I_\alpha^2} + \sigma^2_{I_\beta^2}$$

or

$$\sigma^2_{\Delta I^2} \approx 2\left(\frac{1}{2BT}\right)\sigma^2_{i_\alpha^2}$$

$$= \frac{1}{BT}\, 4i_o^2 \sigma^2_{i_{no}}$$

To obtain the performance of the communication system, one proceeds precisely as in Section 6.4;

$$K = \left(\frac{\overline{\Delta I^2}}{\sigma^2_{\Delta I^2}}\right)^2$$

where K depends on the desired error probability per bit. Thus the information rate H = 1/T becomes

$$H = \frac{B}{K} \frac{i_{so}^2}{\sigma^2_{i_{no}}}$$

or, since i_{no} is a zero mean random variable,

$$H_{horn} = \frac{B}{K} \frac{i_{so}{}^2}{i_{no}{}^2} \tag{39}$$

6.5.2 Discussion

It is instructive to compare this result directly with the heterodyne result. If, for the heterodyne, one makes assumptions comparable to those made for the homodyne. i.e., that i_o, i_s, and i_{so} can be treated deterministically and that the shot noise which affects the decision-making can be limited to a very narrow band, one would obtain [from Equation (36)] that

$$H_{het} = \frac{B}{2K} \frac{\overline{i_{so}^2}}{\overline{i_{no}^2}}$$

Thus, one finds the oft-quoted result that the homodyne enjoys, in principle, a 3 dB advantage over the heterodyne, One should note, however, that for the heterodyne case i_{so}^2 stands for a time-averaged quantity in these analyses. Therefore, since in the heterodyne case i_{so} is an alternating current, the time-averaging introduces an additional factor 1/2, not present in the homodyne, when i_{so} is related to the optical powers and the detector efficiency parameters. Hence in obtaining Equation (37), it was assumed that

$$\overline{i_{so}^2} = 2\left(\frac{\eta\tau_o}{h\nu}\, e\right)^2 P_s P_o = \frac{1}{2} \cdot 4 \left(\frac{\eta\tau_o}{h\nu}\, e\right)^2 P_s P_o$$

where the explicit appearance of the factor 1/2 in this expression comes about from measuring the time average of the alternating IF current i_{so}. In the homodyne case, this factor is not present, so a direct comparison of the two results would appear to indicate that, for the same optical signal power, the homodyne information rate is 6 dB greater than that of the heterodyne.

On the other hand, it is possible in principle to modify the heterodyne receiver so as to measure the energy after converting, a second time, from i-f to dc without incurring any penalty. This would make the heterodyne performance (taking K = 10)

$$H_{het} \approx \frac{1}{10}\eta\tau_o P_s/h\nu$$

In the homodyne case, this "second detection" is not necessary, as a dc term is obtained directly. Thus

$$H_{hom} \approx \frac{1}{5}\eta\tau_o P_s/h\nu$$

Because of similarity to the heterodyne, no illustrative computation of the performance of the homodyne will be carried out. One should remember that the limitations imposed on useful aperture size by limited spatial coherence of the signal beam, and the Doppler shift problem in space communication applications, are common to both receivers.

7. DIRECT DETECTION WITH PMT

Conceptually and in implementation, the simplest receiver for optical communication is that employing direct detection. In this receiver all the energy which lies in a spectral band passed by an optical filter (and which can be focused on an optical detector) is accepted without regard to any phase properties or any spatial restrictions such as apply to mixing. As no special provisions other than the use of the predetection filter are made to discriminate against background noise, this noise will degrade performance of a direct detection system to a greater extent than it will that of a heterodyne receiver. Furthermore, when the incoming signal is weak, as is typical in deep-space communication, the intrinsic detector noises (thermal noise and dark current shot noise) may be large compared to the unavoidable signal shot noise. The detector noises will be prohibitively large unless there is a noiseless (or nearly noiseless) gain mechanism to magnify the signal shot noise (as well as the signal) with respect to the detector noises. In the visible portion of the spectrum, photomultiplier tubes (PMT) exist which can perform this function. In this section the performance of a communication system with binary polarization modulation and direct detection will be analyzed and discussed.

7.1 Sources of Noise

Consider signal plus background radiation falling on a PMT after passing an optical predetection filter and a polarizer. Assume that the signal has the proper polarization so that all the collected optical power in the signal, and half that in the background, strike the PMT. The output current from the PMT is time-averaged over an interval T, assumed long compared to the time constant of the predetection filter. (The integrator acts as a narrow filter of width 1/2T, so no additional filter is included for consideration.)

In addition to the signal contribution and the mean contributions due to background and dark current, the output of the detector will exhibit fluctuations due to a number of sources which are listed and discussed below:

1. Shot noise on the signal
2. Shot noise on the background noise
3. Mixing between signal and background noise
4. Self-beating of background noise*
5. Detector thermal noise
6. Shot noise on detector dark current.

This assumes that the received optical signal does not itself fluctuate.

1. Shot noise on the signal. This is due to the random emission of electrons from the photosurface. It is always present when signal is present. Assuming for the moment a multiplication of unity, when the signal produces an average photocurrent I_s, the variance of the averaged shot noise current (which has zero mean) is

$$\sigma_1^2 = \frac{eI_s}{T}$$

where e is the electron charge.

2. Shot noise on the background noise. This term is also due to the random emission. If the background radiation produces a mean current I_n, the variance of the resulting shot noise current is

$$\sigma_2^2 = \frac{eI_n}{T}$$

Shot noise is experimentally known to have the spectrum of white noise over quite large bandwidths. A more fundamental description of the same phenomenon is given by expressing the output as the number of photoelectrons in some time interval. Then the above relations are equivalent to assuming that the output photoelectrons due to either signal or background are both Poisson-distributed, and hence have variance equal to their mean value. It should be clearly appreciated that the shot-noise behavior is most significant when the number of photoelectrons (or the rate of their production) is relatively small, and when the discrete character of the radiation is most observable.

On the other hand when the degeneracy (the number of photons per mode as discussed in Section 1) of the background radiation is large, the number of quanta in the radiation will be large and the fluctuations will be dominated by the mixing terms introduced as noise sources (3) and (4) above. These can be shown to have the form (Appendix 8):

3. Mixing between signal and background noise.†

$$\sigma_3^2 = \frac{2I_sI_n}{\frac{A_R\Omega_R}{\lambda^2}WT}$$

*In this discussion the self-beating of the background radiation means those contributions which are at baseband, as the others are assumed to vanish in the time integration.

†If the signal occurs in many transverse modes, as will be the case when atmospheric fluctuations spread the received signal out over a large solid angle (with respect to that which the transmitter subtends), then mixing occurs in many modes. However, in each mode the signal contribution to σ_3^2 is consequently reduced and the result mixing fluctuation is the same. That is, for signal in N transverse modes

$$\sigma_3^2 = N \cdot 2\frac{I_s}{N}\,\frac{I_n}{\frac{A_R\Omega_R}{\lambda^2}WT} = 2\,\frac{I_sI_n}{\frac{A_R\Omega_R}{\lambda^2}WT}$$

where A_R, Ω_R are the receiver area and solid angle of view and it is here assumed that the background radiation is isotropic.

4. Self-beating of background noise.

$$\sigma_4^2 = \frac{I_n^2}{\frac{A_R\Omega_R}{\lambda^2}WT}$$

These terms can be significant when compared with σ_1^2, σ_2^2 only if

$$\frac{I_n}{\frac{A_R\Omega_R}{\lambda^2}} \gtrsim eW$$

Since $$I_n = \eta\tau_o \frac{N_\nu}{2} \frac{A_R\Omega_R W}{h\nu} e$$

then $$\frac{I_n}{\frac{A_R\Omega_R}{\lambda^2}} = \eta\tau_o \frac{N_\nu}{2} \frac{\lambda^2}{h\nu} eW = \eta\tau_o \delta eW$$

where η is the quantum efficiency, τ_o the (midband) transmissivity of the optical filter, and δ is the degeneracy of the background radiation (one polarization). Thus for the detection to become essentially classical, i.e., where fluctuations in the wave-like mixing interactions dominate the particle-like shot-noise behavior, the degeneracy δ of the background radiation must be large compared with $(\eta\tau_o)^{-1}$, necessarily larger than unity. This is a situation that never exists at optical frequencies as can be seen from Table 40.

5. Detector or thermal noise. For a detector with effective temperature T_d and resistance R_d, the variance of the (zero mean) thermal current is

$$\sigma_5^2 = \frac{2kT_d}{R_d T}$$

6. Shot noise on detector dark current. This current is due to thermionic emission from the photocathode in the absence of signal and is temperature dependent. As the emission times are random, the fluctuations will also appear as a shot noise current with variance

$$\sigma_6^2 = \frac{eI_d}{T}$$

where I_d is the mean dark current.

The detector fluctuations σ_5^2 and σ_6^2, as mentioned previously, will seriously degrade communication performance unless steps are taken to minimize their importance. In the visible portion of the spectrum this can be done with respect to σ_5^2 using photomultipliers. When the current gain in the PMT is M, then the fluctuations σ_1^2 through σ_4^2 (as well as σ_6^2) are multiplied by M^2. Consider the ratio of background shot noise to thermal noise:

$$\frac{\sigma_2^2}{\sigma_5^2} = \frac{M^2R_d}{2kT_d} eI_n$$

If M^2R_d is sufficiently large, the thermal noise contribution can be neglected. M^2R_d is the power gain, sometimes regarded as a figure of merit for the PMT. In typical deep-space applications, the ratio σ_2^2/σ_5^2 will be large. In such applications, the signal intensity is so weak that it is necessary to use very large receivers (not diffraction-limited) to achieve the desired communication objective. Consider a 10-meter diameter receiver aperture with a field of view of 10^8 steradians and assume there is an optical filter one angstrom wide at a "carrier" of 1 micron. Then for a receiver with $\eta\tau_o = 10^{-2}$ (which may be somewhat conservative) and a background due to sunlight scattered from the atmosphere (Section 1),

$$I_n \approx 10^{-11} \text{ amperes}$$

Thus, assuming a detector temperature $T_d = 300°K$

$$\frac{2kT_d}{eI_n} \approx 5 \times 10^9 \text{ ohms}$$

When M^2R_d exceeds this figure, the thermal noise can be neglected with respect to background shot noise. There are[63] in fact PMT's with power gain upwards to 10^{12}. If instead the background is due to Mars, $2kT_d/eI_n$ will be about four times as large but still small enough with respect to M^2R_d to neglect thermal noise.

It is conceivable that some deep-space missions will be performed with very much smaller background radiation. For example, when the length of the communication link is very much longer, the source of the background radiation subtends a much smaller solid angle and hence I_n will be very small. Even in such a case, the thermal noise σ_5^2 may be small compared to the signal shot noise σ_1^2. Consider

$$\frac{\sigma_1^2}{\sigma_5^2} = \frac{M^2R_d}{2kT_d} eI_s$$

In general, for a given information rate H there is a practical lower bound on I_s. For example, in the binary polarization system considered below,

$$I_s \geqslant 10\,eH$$

Thus
$$\frac{\sigma_i^2}{\sigma_5^2} \geqslant \frac{M^2 R_d \cdot 10e^2}{2kT_d}\,H$$

For a power gain* of 10^{12} ohms and $T_d = 300°K.$,

$$\frac{\sigma_i^2}{\sigma_5^2} \gtrsim 3\text{x}10^{-5}\,H.$$

Thus when H exceeds about 10^5 bits per second, thermal noise will be negligible with respect to signal shot noise. At lower information rates and at very weak backgrounds thermal noise will degrade performance unless carefully cooled PMT's with larger power gain are employed. For the sequel, however, it is reasonable to neglect σ_5^2.

With respect to dark current, this can be neglected with respect to signal shot noise when $\sigma_i^2/\sigma_6^2 = I_s/I_d$ is large. Assuming again that practical systems always operate with more than one signal photoelectron per bit, $I_s \geqslant eH$, one can neglect dark current when $H \geqslant I_d/e$; i.e., when the information rate exceeds the mean rate of emission of dark-current photoelectrons. As the latter can be controlled to small values, less than 1000 per second when cooled, dark current can be neglected at modest information rates.

With the neglect of detector noises and the relative unimportance of the classical fluctuations in optical communications (due to weak degeneracy) one may concentrate solely on the shot noise or quantum regime. The signal-to-noise ratio after time averaging is

$$\text{SNR} = \frac{I_s^2 T}{e(I_s + I_n)}$$

This can also be expressed in terms of the optical power in the signal P_s and in the background P_b:

$$\text{SNR} = \frac{\eta\tau_o}{h\nu}\;\frac{P_s^2\,T}{P_s + \dfrac{P_b}{2}} \qquad (40)$$

7.2 Binary Polarization Modulation

The communication system analyzed is illustrated schematically in Figure 109. The light collector of the receiver has area A_R and solid angle of view Ω_R. Following the optical filter, a polarization separator divides the incoming radiation into the two orthogonal polarizations employed by the transmitter. Each is detected with a PMT. Current at the output is time-averaged for a bit period T and the decision is based on which channel has the larger average current. Perfect synchronization is assumed. Denoting the channel containing the signal by α and the other by β, the observable is $\Delta I = I_\alpha - I_\beta$, which has mean value $\overline{\Delta I} = I_s$. The variance of the difference is

$$\sigma_{\Delta I}^2 = \frac{e}{T}\,(I_s + 2I_n),$$

where the coefficient 2 which multiplies I_n comes from the background shot noise in both channels. Assuming that ΔI has a Gaussian distribution (the validity of this assumption is discussed below), the integration time required to achieve a desired error probability P_ϵ is given by setting

$$K = \frac{(\overline{\Delta I})^2}{\sigma_{\Delta I}^2}$$

where K is determined by the equation

$$P_\epsilon = \Phi(-\sqrt{K})$$

Here P_ϵ is the desired error probability (per bit) and Φ (t) is the (tabulated) function,

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-t^2/2}\,dt$$

For an error probability of 10^{-3}, K (which is quite insensitive to P_ϵ) is approximately 10. Thus, the information rate H which is the inverse of the integration time, is

$$H \approx 0.1\,\frac{I_s^2}{e\,(I_s + 2I_n)} \quad \text{bits/second}$$

In terms of optical powers,

$$H \approx 0.1\;\frac{\eta\tau_o\,\dfrac{P_s}{h\nu}}{1 + P_b\,/\,P_s} \qquad (41)$$

If the background power is distributed in N spatial (transverse) modes, each with the same degeneracy δ, then

$$P_b = 2N\delta h\nu W$$

*For the diode detector as a special type of PMT, the tube impedance at high frequency is principally capacitive. Thus for information bandwidth H, $R_d \approx 1/HC_d$, where C_d is the detector capacitance. For this case, as M = 1,

$$\frac{M^2 R_d \cdot 10e^2 H}{2kT} \approx \frac{e^2}{2kT_dC_d}$$

and thermal noise can be neglected with respect to signal shot noise if this ratio (independent of H) is large.

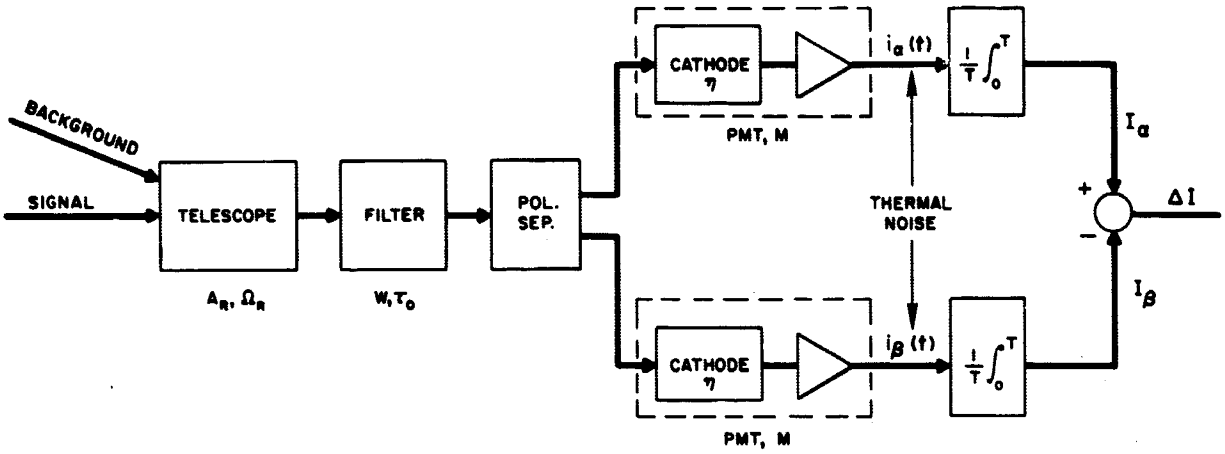


Figure 109. Schematic representation of direct detection receiver

For example, if the background radiation is diffusely spread out over the entire field of view of the receiver, then following Section 1,

$$P_b = 2\frac{\Omega_R A_R}{\lambda^2}\delta h\nu W = N_\nu \Omega_R A_R W$$

where as before, N_ν is power per unit area-solid angle-Hertz, and H can be conveniently expressed in terms of the signal intensity (power per unit area) $\prod_s$ at the receiver:

$$H \approx 0.1 \frac{\eta\tau_o \frac{\prod_s A_R}{h\nu}}{1 + \left(\frac{N_\nu \Omega_R W}{\prod_s}\right)} \qquad (42)$$

There is a further point of interest with reference to the foregoing analysis. The assumption that ΔI is Gaussian, which is justifiable for large WT on the basis of heuristic arguments involving the central limit theorem, may be circumvented. Since one is dealing here with shot-noise limited behavior (on both signal and background), one can carry out the analysis in terms of particle statistics. The number of signal photoelectrons is actually Poisson (assuming a well-stabilized laser is used as the signal source[64]), as is the number produced by the background. Their sum is likewise Poisson and an equation which is almost precisely the same as Equation (42) may be derived directly for a system which compares the number of photoelectrons in the two channels.

7.3 Numerical Examples and Discussion

It is of interest to compute the transmitter power required to achieve certain communication objectives. For greater generality P_b is represented by*

$$P_b = \frac{2A_R W}{\lambda^2} h\nu \sum \delta_i \Omega_i$$

where W is the width in Hz of the predetection filter, Ω_i the solid angle at the telescope which a background noise source i subtends; and i runs over all sources of background noise within the field of view of the receiver. Equation (42) can be rewritten and solved for the required signal intensity at the receiver:

$$\prod_s = \frac{H}{A_R}\frac{10h\nu}{\eta\tau_o}\left\{\frac{1}{2} + \sqrt{\frac{1}{4} + 2\frac{A_R W \sum \delta_i \Omega_i}{\lambda^2}\frac{\eta\tau_o}{10H}}\right\} \qquad (43)$$

The signal intensity can be expressed as

$$\prod_s \approx \tau_A P_t \left(\frac{D_t}{R_o \lambda}\right)^2$$

where τ_A is the attenuation due to the atmosphere, P_t the transmitter power, D_t the diameter of the transmitter, and R_o the length of the communication path. Equation (43) has been used to calculate P_t as a function of the desired information rate, the results of which are plotted in Figures 110 through 113. In these calculations it has been assumed that the predetection filter has a width of one angstrom, $\eta\tau_o = 10^{-2}$ (corresponding, for example, to $\eta = 0.05$ and $\tau_o = 0.2$), $D_t = 1$ meter, and $R_o = 1.6 \times 10^{11}$ meters (10^8 miles). The communication is assumed to come from a vehicle in the vicinity of Mars and, for a satellite receiver, sunlight scattered from Mars is taken as representative of the background radiation. In the calculation for a receiver on the ground, where the telescope has been taken to have $\Omega = 10^{-8}$ steradian, corresponding to the atmosphere induced broadening of the received beam,[65] the contribution due to atmosphere-scattered sunlight is about one order to magnitude greater than that due to scattering from Mars and the latter has been neglected. For the ground receiver, the atmospheric transmissivity $\tau_A = 0.7$ has been assumed[65] which corresponds to good seeing conditions.

An inspection of Figures 110 through 113 shows that, for small information rates, a large additional power is required to overcome the effect of background noise, i.e., relative to the power required at what is usually called the shot noise limit:

$$\prod_s = \frac{H}{A_R}\frac{10h\nu}{\eta\tau_o}$$

(At the shot noise limit, 10 photoelectrons are required for this system for each bit of information.) When the information rate is large, however, the additional power required over the shot-noise limit is small. An inspection of Equation (43) shows that the effect of background noise (which leads to requiring additional power) is contained in the size (relative to unity) of the term

$$\frac{1}{10}\frac{\frac{(\eta\tau_o)2A_R\left(\sum \delta_i \Omega_i\right) W}{\lambda^2}}{H}$$

*A familiar illustration of the use of this relation is furnished by considering a thermal source at low frequencies, $h\nu \ll kT$, being viewed by a diffraction-limited receiver. Then $\lambda^2/A_R = \Omega_R$, $\delta = kT/h\nu$ (for one polarization), $\Omega = \Omega_{\text{source}}$ ($< \Omega_R$). Hence $P_b = kTW\,\Omega_{\text{source}}/\Omega_R$.

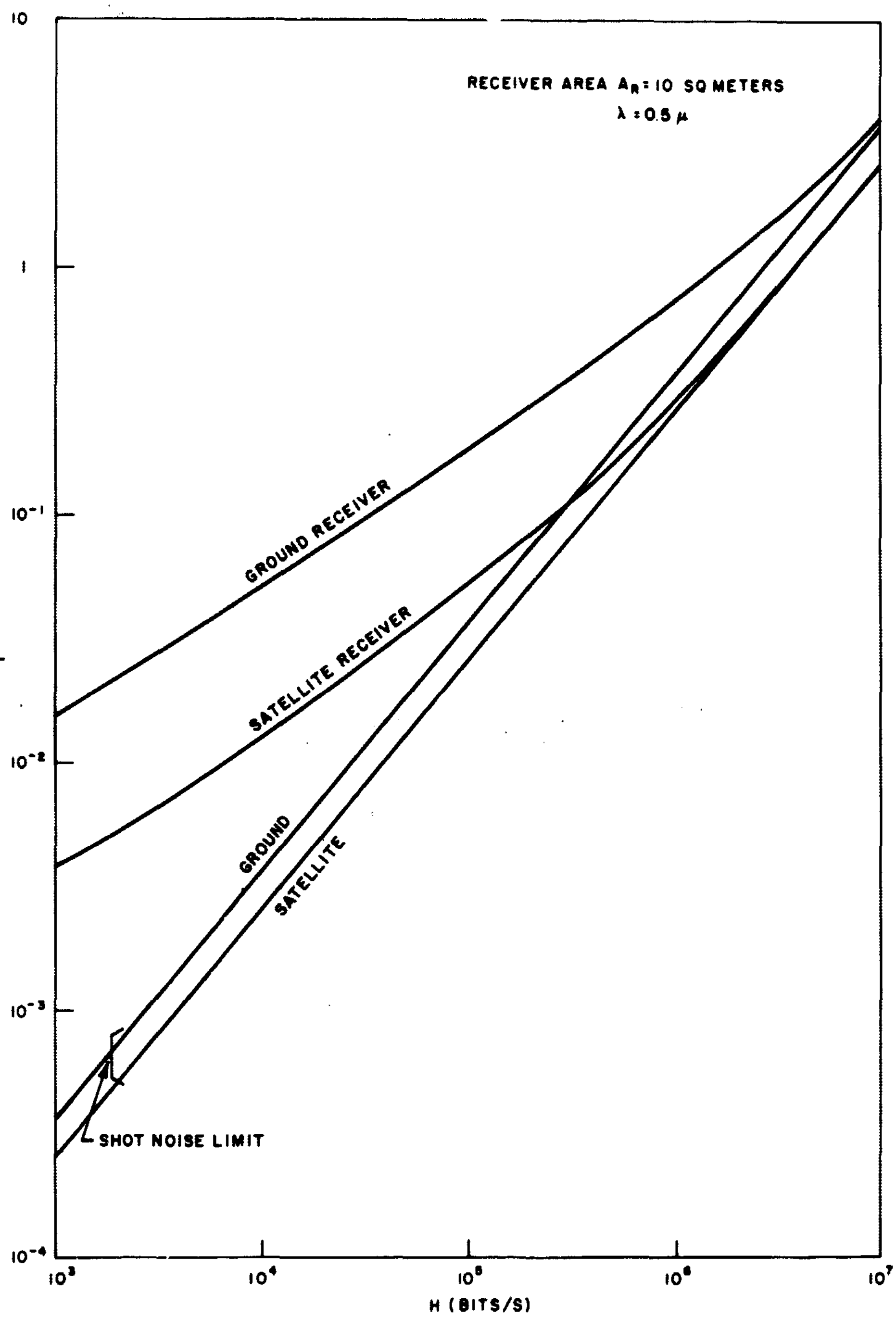


Figure 110. Direct detection receiver: transmitter power vs. information rate

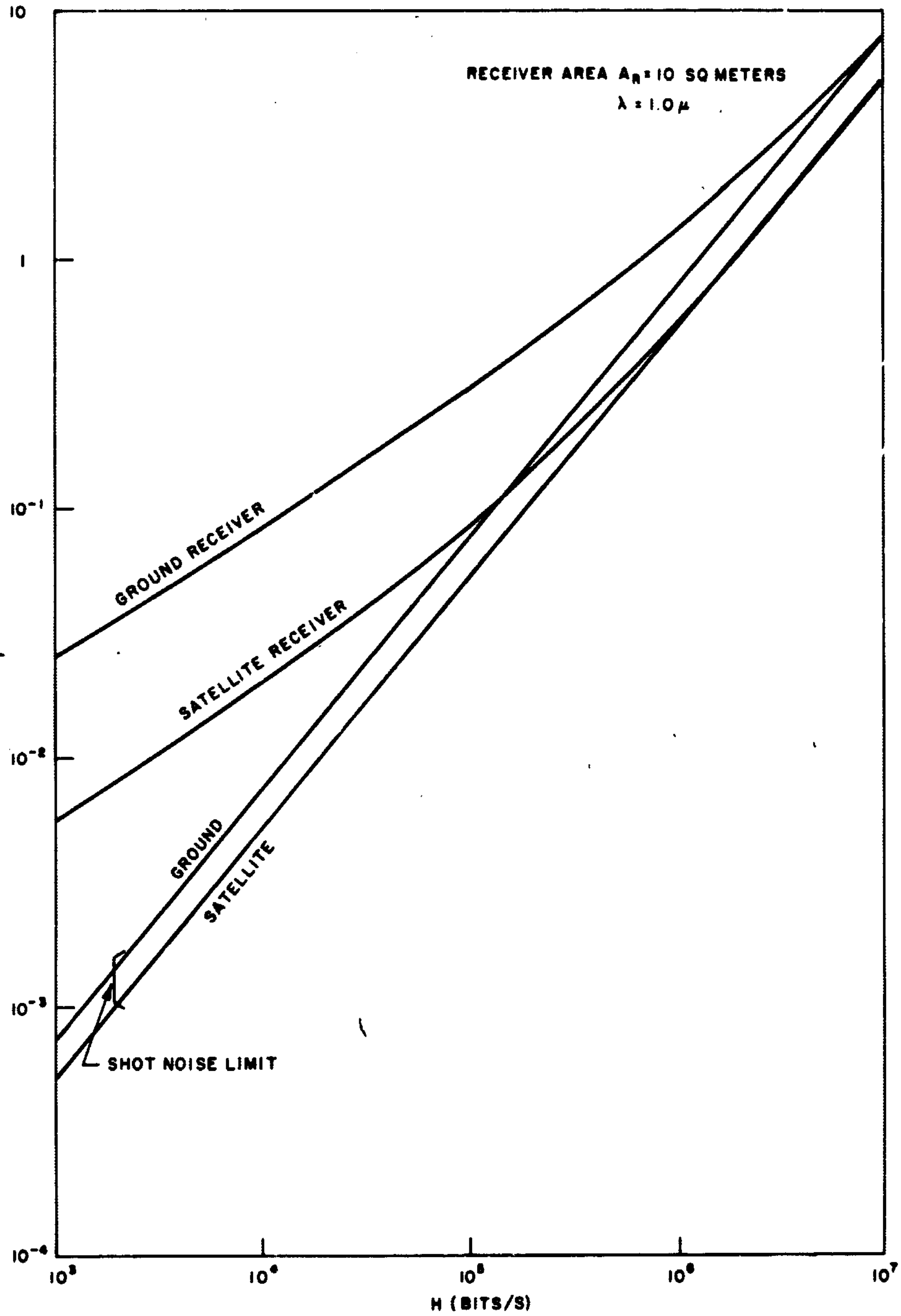


Figure 111. Direct detection receiver: transmitter power vs. information rate

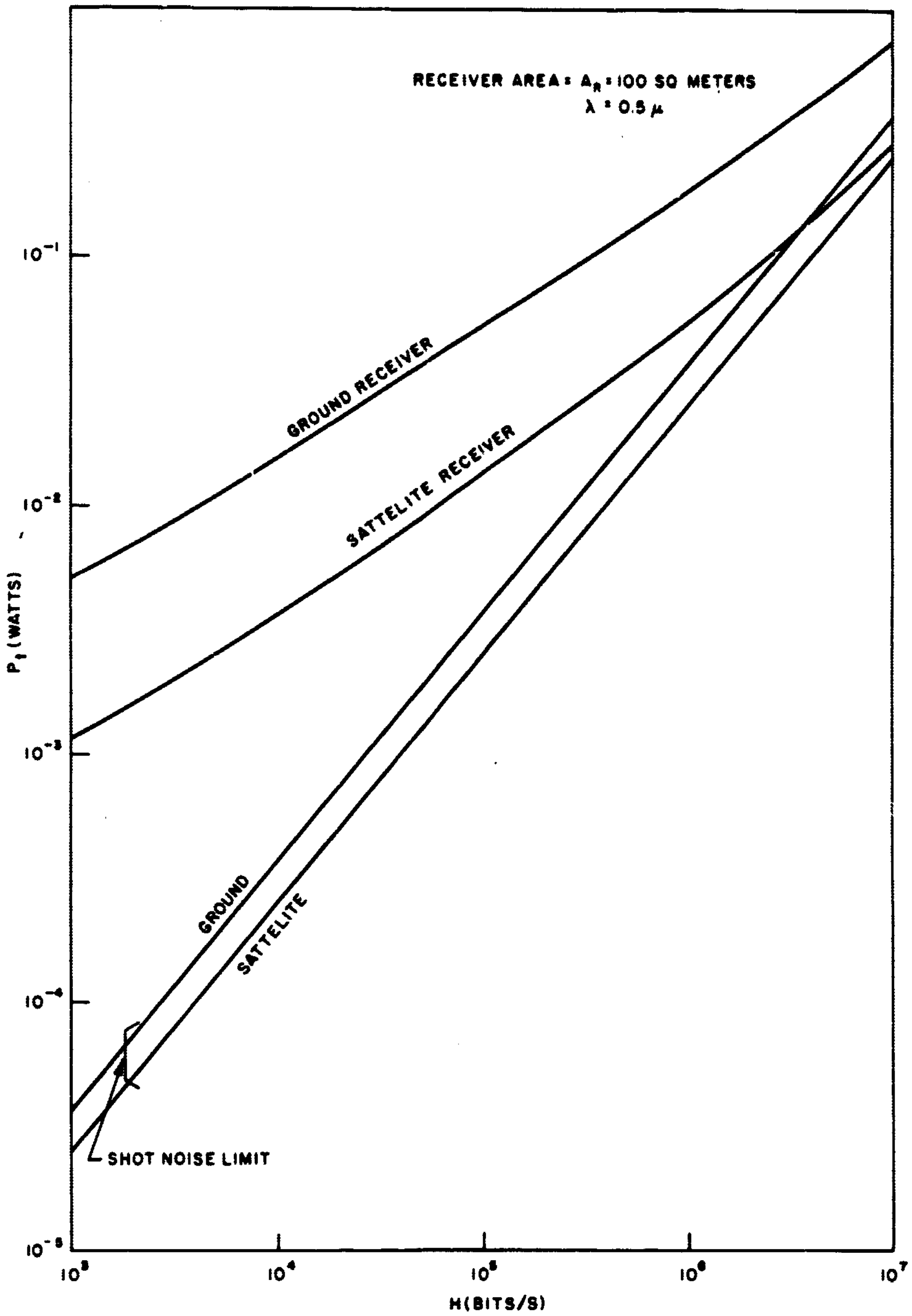


Figure 112. Direct detection receiver: transmitter power vs. information rate

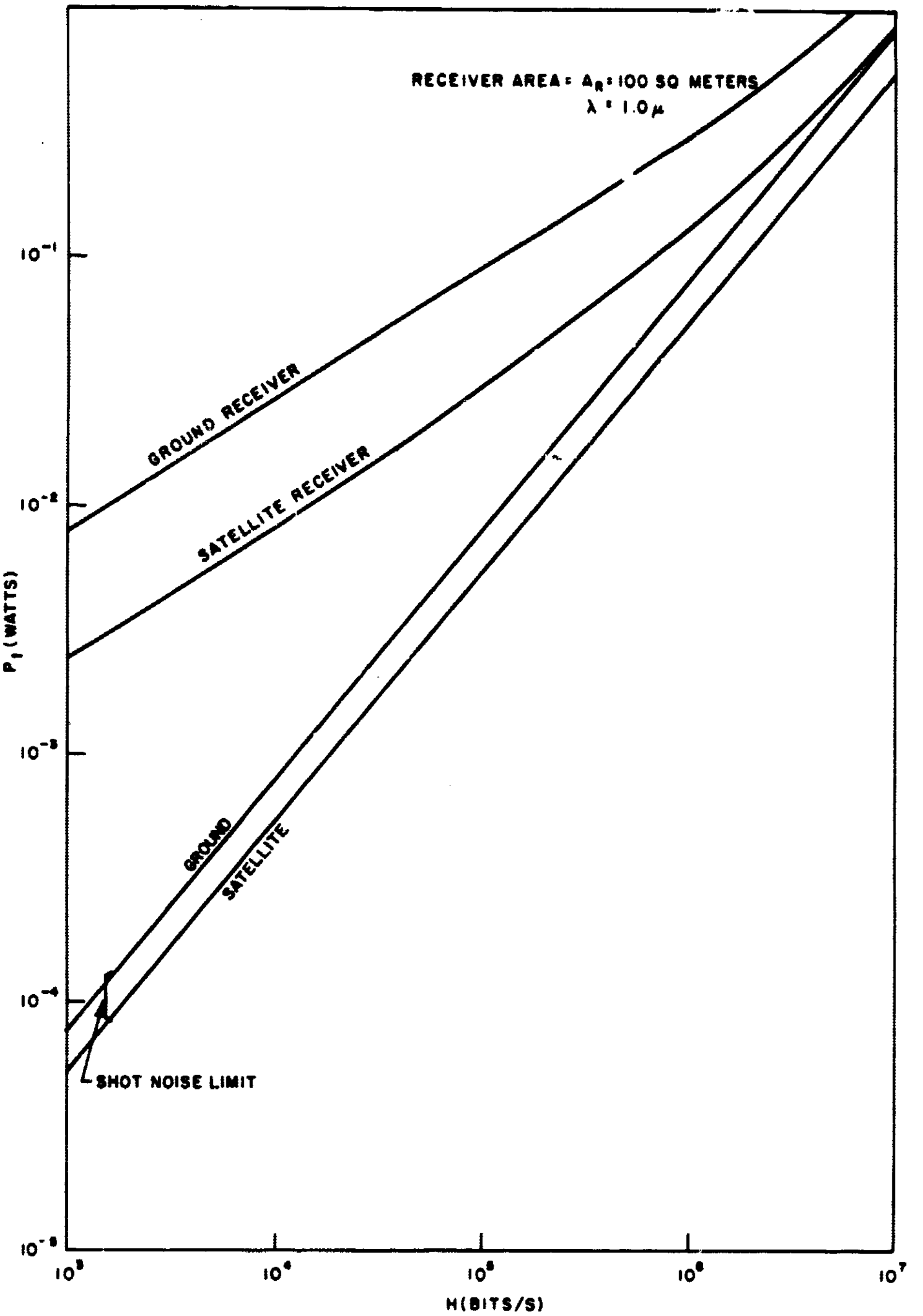


Figure 113. Direct detection receiver: transmitter power vs. information rate

This term can be readily seen to be 1/10 the ratio of the rate at which photoelectrons due to the background (both polarizations) are generated to the information rate. Hence, when the desired information rate exceeds approximately 1/10 the rate of photoelectron generation from background noise, the shot-noise limit is virtually achieved (since then the rate of photoelectron generation from the signal must be correspondingly strong). At an information rate of 10^6 bits per second, the ratio in question is not large ior a receiver of 10 square meters. This small a receiver may only be used if there is adequate transmitter power available, about one watt or more (for a distance of 1 AU). For an available transmitter power of approximately 1/10 watt, a larger receiver must be used, the ratio of background photoelectrons to information rate is larger, and there is a more significant degradation due to the background.

The term which implies the degradation due to background noise is proportional to the filter width W, hence the required power grows only as $W^{1/2}$ even in the background-limited regime.

In the foregoing discussion, it has been assumed that all the collected signal power is detected, hence that the photodetector is large enough and the telescope optics constructed and maintained with sufficient accuracy to make this a reality. That is, all the spatial modes of the signal are collected and it is assumed that intensity fading can be neglected. Finally, it should be observed that the angular divergence of the transmitted beam has implicitly been assumed to be 1/2 microradian (at 0.5 μ wavelength) so that a stringent requirement for pointing the transmitter and the receiver has been imposed. The validity of each of these assumptions is open to question and is discussed elsewhere in various parts of this report.

8. OPTICAL PREAMPLIFIER

The optical preamplifier receiver furnishes another method of enhancing the signal and some of the noise fluctuation with respect to the detector noises. In such a system the preamplifier makes the rate at which photons strike a photodetector so large that the particle-like shot noises become negligible compared with the wave-like mixing noises. As a result, the performance of such a system is independent of quantum efficiency and the transmission losses in the optics as will be shown.

8.1 Analysis

The signal-to-noise performance of an optical preamplifier receiver can be most easily understood by regarding the receiver as a variation of a direct-detection receiver. Assume that a polarized signal plus background noise of the same polarization are passed through an optical filter of width W and subsequently amplified with gain G before the output of the amplifier strikes a photodetector with quantum efficiency η. The output current is subsequently time-averaged over an interval T. Assume that the input to the amplifier would for unit gain produce an average signal current I_s and an average current due to the background radiation I_n. Then the fluctuation terms listed in the analyses of the direct detection system, Section 7.1, would be modified by a factor G for each current contribution which appears in that section. Consequently, listing the variances of the time-averaged currents:

Signal shot noise: $\sigma_1^2 = GeI_s/T$

Shot noise on background: $\sigma_2^2 = GeI_n/T$

Mixing between signal and background noise:

$$\sigma_3^2 = G^2 \cdot \frac{2I_s I_n}{MWT}$$

Self-beating of background noise: $\sigma_4^2 = G^2 \cdot \frac{I_n^2}{MWT}$

In writing these variances, it has been assumed that the signal is evenly spread into M spatial modes, where M is the ratio of receiver area to the coherence area of the signal. It is assumed that the amplifier is designed to provide gain G for these transverse modes but none other and all M modes are seen by the detector.[65] There is another important difference between this analysis and that of the direct detection receiver previously considered; namely, that the spontaneous emission noise of the amplifier is added to the background and amplified. This amounts ideally to one photon per Hz for each mode.[66]
Hence

$$I_n = \eta\tau_o MW(1+\delta)e$$

where δ is in number of photons per Hz per transverse mode in the background radiation and τ_o is the transmissivity. Since at optical frequencies $\delta << 1$ (Section 1)

$$I_n \approx \eta\tau_o MWe$$

As before

$$I_s = \eta\tau_o \frac{P_s}{h\nu} e$$

Comparing the particle-like terms with the mixing terms,

$$\frac{\sigma_2^2}{\sigma_4^2} = 2\frac{\sigma_1^2}{\sigma_3^2} = \frac{eMW}{GI_n} = \frac{1}{G\eta\tau_o}$$

Hence if $G >> (\eta\tau_o)^{-1}$, the shot-noise terms can be neglected.

In addition to the above fluctuations one has, once again, tl.e detector noises:

$$\text{Thermal noise:} \qquad \sigma_5^2 = \frac{2kT_d}{R_d T}$$

$$\text{Dark current shot noise:} \qquad \sigma_6^2 = \frac{eI_d}{T}$$

Comparing thermal noise with the background self-beating fluctuations (which are always important for this receiver because of the amplified spontaneous emission noise) gives

$$\frac{\sigma_4^2}{\sigma_5^2} = \frac{(G\eta\tau_o)^2}{2kT_d/MWR_d e^2}$$

Hence, if

$$(G\eta\tau_o) \gtrsim \left[\frac{2kT_d}{MWR_d e^2}\right]^{1/2}$$

thermal noise may be neglected.* It should be noted that the predetection filtering W may be imposed by the amplifier gain characteristics (e.g., the width of the laser transition with Doppler broadening). For a CO_2 laser amplifier (10.6μ), this is about 0.5×10^8 Hz. Assuming this bandwidth and $T_d = 300$ degrees, $R_d = 10^2$ ohms, and $M = 1$, one gets

$$\left[\frac{2kT_d}{WR_d e^2}\right]^{1/2} \approx 7\text{x}10^3$$

However, the equivalent resistance may be quite a bit higher and, for 10.6μ,† M will be larger as well. Typically, M may be expected to range between 10 and 100, while R_d may range from 10^4 to 10^7 ohms. One is justified in anticipating, therefore, that amplifier gain of perhaps 30 to 40 dB will be more than adequate so that the thermal noise will be negligible with respect to background self-beating noise for this system. This being the case, one may also expect $G >> (\eta\tau_o)^{-1}$, hence the shot noise contributions will also be negligible compared with the mixing noises.

The ratio of noise self-beating to detector dark current must also be examined. For a photoconductive detector at 10.6u. the dark current is likely to be the most significant detector noise. (As discussed in Section 6.2, in rating such detectors the dark current fluctuation, principally due to generation-recombination, is lumped with other noise sources.) The ratio is easily seen to be

$$\frac{\sigma_4^2}{\sigma_6^2} = (G\eta\tau_o)^2 \frac{MWe}{I_d}$$

Hence if

$$G\eta\tau_o > \left(\frac{I_d/e}{MW}\right)^{1/2}$$

the dark current is negligible. For a good detector cooled to low temperature, I_d/e is of the order 10^{16} photoelectrons per second. Thus for M = 50 and W = 50 MHz, one can neglect dark current if

$$G\eta\tau_o \gtrsim 0.2 \times 10^4$$

This implies, taking $\eta\tau_o = 0.2$, a gain of about 40 dB.

Summarizing, one finds that for gain somewhat exceeding 40 dB (and when great care is taken to reduce detector noise), one can expect the dominant fluctuation terms to be $\sigma_3{}^2$, $\sigma_4{}^2$, the wave-like effects due to mixing. The resulting signal-to-noise ratio becomes

$$\text{SNR} = \frac{G^2 I_s^2}{G^2 \cdot \frac{2I_s I_n}{MWT} + \frac{G^2 I_n^2}{MWT}} = \frac{1}{2} \frac{\frac{I_s}{I_n/MW}}{1 + \frac{I_n}{2I_s}} T$$

In terms of the total optical power P_s (in one polarization)

$$\text{SNR} = \frac{1}{2} \frac{P_s/h\nu}{1 + \frac{Mh\nu W}{2P_s}} \qquad (63)$$

Note that this result is independent of detector quantum efficiency η and transmissivity of the optics τ_o. It is important to observe that P_s is the signal intensity integrated over the entire optical aperture: any limitation imposed on the signal-to-noise ratio by the limited coherence of the signal over the aperture is in the quantity $M = A_R/a_s$ (>1), where a_s is the coherence area of the signal.

*Note that at strong signal power, when signal-background mixing exceeds the self-beating, this criterion for neglecting thermal noise may be unduly stringent.

†Emphasis is here placed on a preamplifier system for the infrared laser. In the visible region, high gain PMT's will by themselves limit the importance of thermal noise. Furthermore (as will be subsequently shown), at visible frequencies direct detection with a PMT should give superior performance to a system employing an optical preamplifier.

8.2 System Performance

Consider a simple communication system using binary polarization shift modulation which is a straightforward modification of that analyzed in Section 7. The only difference is that the polarized signal plus background radiation is passed through the optical preamplifier before it strikes the photodetector.* As before, the decision as to the polarization of the incoming signal is made on the basis of the larger time-averaged current. The difference in the two time-averaged currents contains two contributions to the self-beating. Thus, if the difference is denoted by ΔI,

$$\sigma^2_{\Delta I} = G^2 \left\{ \frac{2I_s I_n}{MWT} + \frac{2I_n^2}{MWT} \right\}$$

Repeating the previous procedure and assuming ΔI is Gaussian, for $P_\epsilon = 10^{-3}$ the information rate becomes:

$$H \approx 0.05 \frac{P_s/h\nu}{1 + \frac{Mh\nu W}{P_s}} \qquad (64)$$

or

$$H \approx 0.05 \frac{(\pi_s/h\nu)A_R}{1 + \frac{h\nu W}{\pi_s \min(A_R, a_s)}} \qquad (65)$$

Note that Equation (65) does not exhibit a saturation of information rate with receiver area (as does the heterodyne receiver) and for receivers larger than the coherence size H is simply linear with receiver area.

In Figure 114 is plotted the transmitter power required to achieve a desired information rate for a 10 micron transmitter: It is assumed that atmospheric transmission is 0.7, the coherence area is 1 square meter, $W = 10^8$ Hz, the length of the communication link 10^8 miles, and the receiver area 10 to 100 square meters ($M = 10, 100$). It can be observed that over most of these curves the transmitter power required is inversely proportional to the square root of the receiver area. This can be seen from Equation (65): For all but the largest information rates, the denominator of Equation (65) is dominated by the amplifier noise term, thus (since $A_R > a_s$) the information rate goes as $\pi_s^2 A_R$.

For space communication, a complication is added to the amplifier receiver (in common with the heterodyne receiver). The bandwidth of a 10.6μ amplifier is about 50 MHz, which is much too narrow to accommodate the changes of frequency of the received light resulting from Doppler effects. For some amplifiers this difficulty could be overcome by tuning the amplifier response to correspond to the frequency of the incoming light. The 10.6μ CO_2 amplifier is not tunable, but this does not rule out the possibility of using it. At the receiver, the incoming signal could be translated in frequency to be always in the passband of the amplifier. This will result in some loss of signal power.

9. COMPARISON OF RECEIVERS

In this section, partly for the sake of summary, the idealized performance of various types of receivers considered in the previous sections will be compared: heterodyne, direct detection, and (optical) preamplifier receivers.

9.1 Visible Frequencies (0.5μ)

9.1.1 Direct Vs. Heterodyne Detection

The very serious shortcoming of a heterodyne receiver situated on the ground is the restriction on useful receiver aperture size imposed by atmospheric transmission. This will be more than enough to counterbalance, in favor of direct detection, the possibility of achieving shot-noise-limited performance for the heterodyne receiver. For a receiver on a satellite, where this restriction on aperture size is not applicable, the heterodyne may enjoy an advantage. However, as will be shown, one can anticipate this advantage will be slight because the background environment is usually less severe for a satellite than a ground receiver and the direct-detection receiver itself may operate close to the shot-noise limit when on a satellite.

Consider the expressions previously derived for information rate, letting the subscripts H and D denote heterodyne and direct respectively. Assuming that the heterodyne employs double detection

$$H_H = \frac{\eta\tau_o}{10} P_{SH}/h\nu$$

$$H_D = \frac{\eta\tau_o}{10} \frac{P_{SD}/h\nu}{1 + P_b/P_{SD}}$$

Setting $H_H = H_D = H$, the desired information rate, and solving for the ratio of signal power required in the direct case to the heterodyne case, one readily finds

$$\frac{P_{SD}}{P_{SH}} = \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\eta\tau_o P_b}{10\,h\nu H}}$$

*Note that care must be taken to see that in each channel, which contains radiation of a single polarization, the amplified spontaneous emission noise is again polarized. If this is not done, there will be self-beating contributions from each of two polarizations which add extraneous noise. In practice it should be possible to eliminate the unwanted polarization of the noise in each channel.

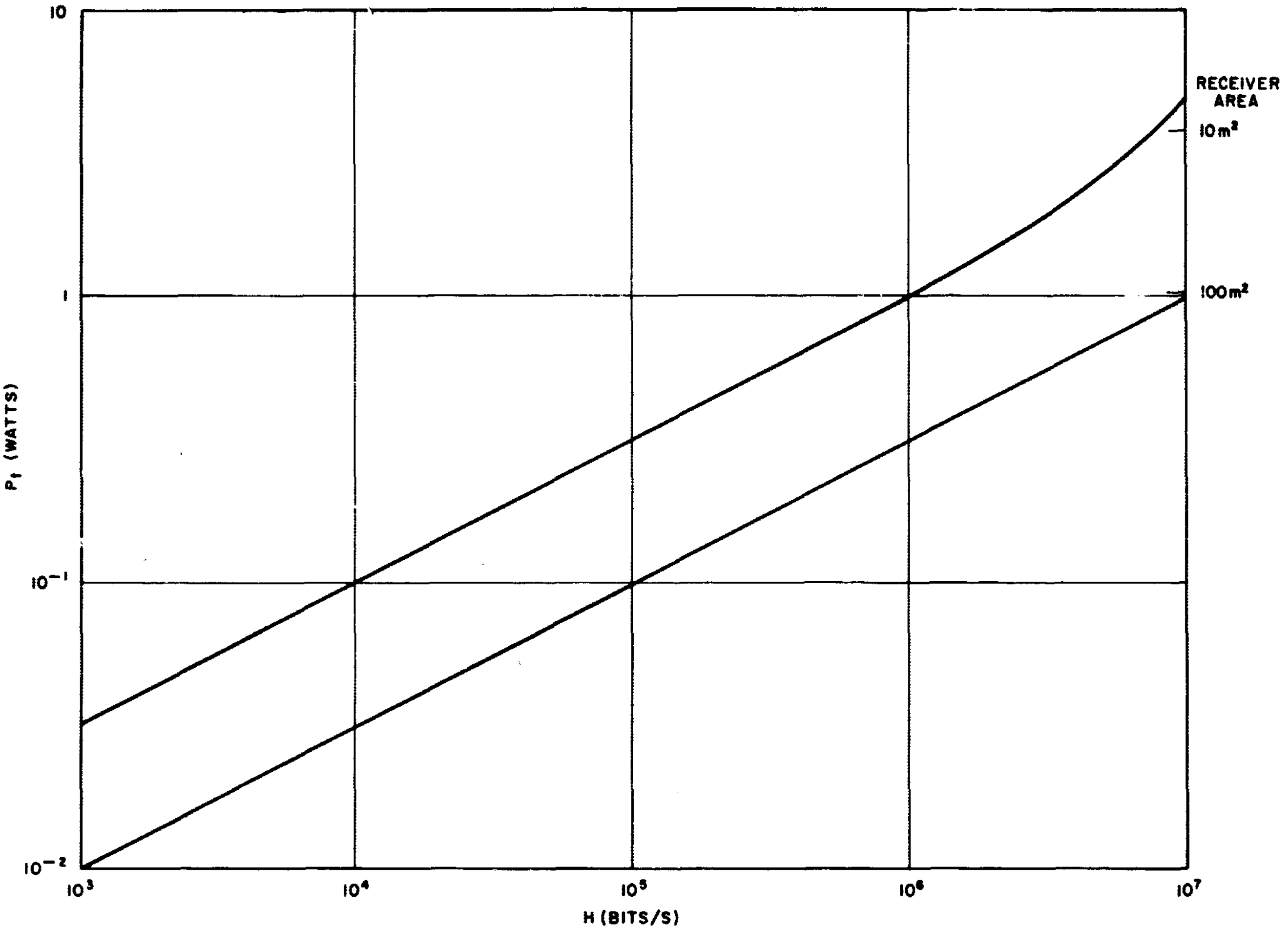


Figure 114. Optical preamplifier receiver: transmitter power vs. information rate (λ = 10μ)

As was done previously, P_b may be expressed in terms of the degeneracy δ as well as the number of spatial modes of background noise seen by the receiver, $N = A_R\,\Omega/\lambda^2$

$$P_b = (2\delta)\ N \cdot W \cdot h\nu$$

Writing the above equation in terms of the receiving aperture and transmitter powers,

$$\frac{P_{tD}}{P_{tH}} = \frac{a_s}{A_R}\left[\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\eta\tau_o NW\delta}{5H}}\right] \qquad (67)$$

where, as before, a_s is the signal coherence area. Equation (67) can also be recast in terms of A_R/a_s as follows:

$$\frac{P_{tD}}{P_{tH}} = \left(\frac{a_s}{A_R}\right)\left[\frac{1}{2} + \sqrt{\frac{1}{4} + \left(\frac{a_s}{a^*}\right)\left(\frac{A_R}{a_s}\right)}\right]$$

where

$$a^* \equiv \frac{5H\lambda^2}{(\eta\tau_o)W\delta\Omega}$$

In this form, it is readily seen that P_{tD}/P_{tH} is a monotonic decreasing function of A_R/a_s. The latter, in turn, has a minimum value of 1, so that

$$\left(\frac{P_{tD}}{P_{tH}}\right)_{max} = \frac{1}{2} + \sqrt{\frac{1}{4} + \left(\frac{a_s}{a^*}\right)}$$

and is a function of H, tending toward unity when H is large. In fact, one can anticipate that in practice, for $H = 10^6$ bits per second, a_s/a^* is a small quantity. Let $\lambda = 0.5\mu$, $\eta\tau_o = 5\text{x}10^{-2}$, $W = 6\text{x}10^{10}$ Hz (0.5 angstrom optical filter), $\delta = 2\text{x}10^{-8}$ and let Ω be a figure representative of limiting of the aperture by the atmospheric turbulence, $\Omega_R = 10^{-8}$. Then $a^* \approx 2$ square meters, whereas a_s will be much smaller than this (Section 2). Thus the maximum value of P_{tD}/P_{tH} (on the ground) for $H = 10^6$ will be very close to unity and more generally, for $A_R >> a_s$, the direct-detection system will require much less power than the heterodyne. For example, assuming (in addition to the foregoing parameters) that $A_R = 25\ m^2$ (5 meter aperture) and $a_s = 25\text{x}10^{-4}\ m^2$ (5 cm coherence length),

$$P_{tD}/P_{tH} \approx 4\times10^{-4}$$

The heterodyne system requires about 34 dB more transmitter power to achieve an information rate of 10^6 bits/second on the ground than a direct-detection receiver for 0.5μ radiation.

The foregoing discussion applies to a receiver on the ground. For a receiver on a satellite, it may be assumed that $a_s = A_R$, but Ω will likely be smaller than for a ground receiver as it is the angle subtended by the source of the background radiation. Considering the background to be Mars at 10^8 miles and again letting $H = 10^6$ bits/second, one finds

$$\frac{P_{tD}}{P_{tH}} \approx 1.6$$

That is, even though the coherence area is not limiting, the background is a less severe problem so that only a 2 dB advantage is obtained by using a heterodyne receiver.† The inherent simplicity of direct detection more than outweighs this small potential advantage. Thus, direct detection will be a clear choice over heterodyne detection at visible frequencies.

9.1.2 Direct Detection Vs. Preamplifier Receiver

The potential advantage of the preamplifier receiver over direct detection is that its performance is independent of the detector efficiency if the signal power is adequate to overcome the spontaneous emission noise that is introduced by the amplifier. In the case of visible frequencies, the noise introduced in the amplifier is normally so large that one cannot achieve this potential advantage.

Let the subscript P denote the optical preamplifier system and compute the ratio of transmitter powers required for each of the receivers for (equal) information rate H. Let M denote the number of spatial modes in which the signal is found ($M = A_R/a_s$) and which the receiver is ideally designed to amplify. One finds, after some straightforward algebra:

$$\frac{P_{tP}}{P_{tD}} = 2(\eta\tau_o)\left|\frac{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{MW_P}{20H}}}{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{(\eta\tau_o)NW_D\delta}{5H}}}\right| \qquad (68)$$

In this expression, W_P is the gain-bandwidth of the amplifier and W_D the width of the optical predetection filter. Since $M \geqslant 1$, the quantity $MW_P/20H$ will, for information rates that are of interest (say, $H = 10^6$), dominate the numerator. For example, an estimate of W_P may be several GHz, in which case $MW_P/20H \geqslant 10^2$. That is, as indicated above, one is usually not operating the preamplifier receiver in the "shot-noise limit." On the other hand, while W_D will usually exceed W_P, it may be only

† In certain applications (missions) the field of view of the receiver may not include the planetary background at all. Then both receivers can be shot-noise limited and hence will require equal power.

about one order of magnitude larger, so that $(\eta\tau_0)MW_D\delta/5H$ is much smaller than $MW_P/20H$. Physically, this simply says that the amplifier is producing many more noise photoelectrons than would be caused by the natural background light. Only when H is very large can the ratio P_{tP}/P_{tD} become unity, and it must be larger yet before the ratio can approach its asymptotic value, $2(\eta\tau_0)$.

Therefore, from a theoretical performance viewpoint, direct detection is preferable to use of a preamplifier at visible frequencies. From a practical viewpoint, the difficulty of achieving appreciable gain at visible frequencies may be an even more compelling reason for using direct detection. This is despite the fact that, since high-gain PMT's may be used in conjunction with an optical preamplifier, the amplifier gain required is not large. That is, to obtain the limiting performance of Equation (64), it is only necessary at visible frequencies that the amplifier gain $G>>1/\eta\tau_0$, since the gain in the PMT itself can be used to overcome the inherent detector noises.

9.2 Infrared Frequencies (10μ)

9.2.1 Heterodyne Vs. Preamplifier Detection

At this frequency, direct detection need not be considered as there are no detectors with sufficient internal gain such that detector noises become negligible.

To compare preamplification with heterodyne one can, after straightforward manipulation, write the ratio of transmitter powers required:

$$\frac{P_{tP}}{P_{tH}} = \frac{\eta\tau_0}{M}\left\{1 + \sqrt{1 + \frac{MW_P}{5H}}\right\} \qquad (69)$$

As H becomes large, the asymptotic ratio becomes $2(\eta\tau_0)/M$ which, for a receiver on the ground where N might be large, is quite a significant advantage to the preamplifier receiver. However, taking $W_P = 5\times10^7$ Hz, $MW_P/5 = 10^7 M\ \mathrm{sec}^{-1}$ so that only if H is unrealistically large may this great a benefit be anticipated. More typically, for $H = 10^6$, one has

$$\frac{P_{tP}}{P_{tH}} \approx \eta\tau_0 \sqrt{\frac{W_P}{5MH}}$$

For $H = 10^6$, $W_P = 5\times10^7$ Hz, $P_{tP}/P_{tH} \sim \eta\tau_0\sqrt{10/M}$. For large M, this can still give an advantage to the preamplifier system. For example, suppose the coherence length for 10μ radiation is one meter, then for a 10 meter diameter receiver, $M = 10^2$. Assuming a "detector efficiency," $\eta\tau_0$, of 0.2, the heterodyne system requires about 12 dB more transmitter power than the preamplifier system. Actually, the above relations (as well as the analysis in Section 8) assume an ideal amplifier. In reality there is not one photon per mode in the effective input but

$$\frac{1}{\epsilon} = \frac{n_2}{n_2 - n_1}$$

photons per mode, where n_1, n_2 are the occupation numbers for lower and upper states of the laser. More accurately, then, it can be easily shown that

$$\frac{P_{tP}}{P_{tH}} \approx \frac{\eta\tau_0}{\epsilon}\sqrt{\frac{10}{M}} \text{ (for H = } 10^6 \text{ bits per second)}$$

Hence, for less optimistic assumptions about both the receiver antenna and the amplifier efficiency,* the power advantage to the preamplifier system might diminish somewhat. For example, if $\epsilon = 1/2$ and the receiver aperture is 5 meters, the advantage shrinks to about 5 dB.

For a receiver on a satellite the situation is different because $M = 1$. In this case the preamplifier system may require more transmitter power than the heterodyne.

Thus, for a 10μ communication system, the preamplifier receiver is slightly more efficient than the heterodyne for a ground receiver and comparable or slightly inferior for a satellite receiver.

9.3 Visible Vs. Infrared Performance

Having investigated, insofar as possible under assumptions it was necessary to make, the best ideal systems for visible and infrared radiation for both ground and satellite receivers, it is now desirable to compare the performance of these best systems relative to one another. To do so, calculate the ratio of transmitter powers required to achieve an information rate H.

9.3.1 Ground Receiver

Here a direct detection receiver (D) at $\lambda_D = 0.5\,\mu$ is compared with a preamplifier receiver (P) at $\lambda_P = 10\,\mu$. The

*Under certain circumstances, the efficiency ϵ of a CO_2 laser will approach 1, since the lower level depopulates much more rapidly after pumping than the upper level. If the principal mechanism for depopulating the upper level is relaxation, then ϵ may be as large as 0.9.

ratio of transmitter powers P_t required is calculated from above relations to be:

$$\frac{P_t(10\mu)}{P_t(0.5\mu)} = \frac{\lambda_D\ A_{RD}}{\lambda_P\ A_{RP}} \frac{2(\eta\tau_o)_D}{\epsilon}$$

$$\left[\frac{1 + \sqrt{1 + \frac{MW_P}{5H}}}{1 + \sqrt{1 + \frac{4(\eta\tau_o)_D \cdot \delta \cdot NW_D}{5H}}}\right]$$

In this equation it has been assumed that the transmitter apertures are the same for both cases; i.e., that it is not significantly easier to make a sizable (of the order of a meter) antenna for $10\,\mu$ than for 0.5 micron when the antenna must perform the transmitting function from space. Assuming now that the receiver apertures are likewise the same and that $\epsilon = 1$,

$$\frac{P_t(10\mu)}{P_t(0.5\mu)} = 40(\eta\tau_o)_D \left[\frac{1 + \sqrt{1 + \frac{MW_P}{5H}}}{1 + \sqrt{1 + \frac{4(\eta\tau_o)_D NW_D\delta}{5H}}}\right]$$

Consider again a calculation for $H = 10^6$ bits/second. Taking $W_P = 50$ MHz, $MW_p/5H$ dominates unity. Let $N = A_R\ \Omega/\lambda^2$ (where $\lambda = 0.5\mu$) and note that for representative figures, taking a large receiver, the quantity $4(\eta\tau_o)_D NW_p\delta/5H$ will also dominate unity. (For example: $A_R = 25\ m^2$, $\Omega = 10^{-8}$, an estimate of the atmosphere imposed limit, $W_D = 6 \times 10^{10}$ Hz, $H = 10^6$, $(\eta\tau_o)_D = 5 \times 10^{-2}$, $\delta = 2 \times 10^{-8}$ gives a value 48.)

Thus

$$\frac{P_t(10\mu)}{P_t(0.5\mu)} \approx 20 \sqrt{\frac{(\eta\tau_o)_D \lambda_D^2 W_P}{a_{SP}\ \Omega\delta W_D}}$$

$$\approx 20\ \frac{\lambda_D}{d_{SP}} \sqrt{\frac{(\eta\tau_o)_D W_P}{\Omega\delta W_D}}$$

Using the above values gives

$$\frac{P_t\,(10\mu)}{P_t\,(0.5\mu)} \approx \frac{4.5}{d_{SP}}$$

where d_{SP} is the coherence length for $10\,\mu$ radiation in meters. If this is estimated to be one meter, this gives a factor of 4.5 more transmitter power required for the 10μ system than for the 0.5μ system. On the other hand, the 10 micron transmitter is estimated to have greater overall power efficiency (see Chapter 4, Section 1.2) than the $0.5\,\mu$. While precise quantitative estimates are premature, it appears likely that the 10μ system will require less overall power in the space vehicle to achieve 10^6 bits per second information rate to a ground receiver.

The absolute values of transmitter powers required are compared in the following table.

Range (mi)	P_t (0.5μ) (watts)	P_t (10μ) (watts)
10^8	0.23	1.0
10^9	23	100

The assumptions upon which these results are based are given in the following table.

Assumptions

	0.5μ	10μ
Receiver area, A_R	25 m^2	25 m^2
Transmitter diameter, D_t	1 m	1 m
Detector efficiency, $\eta\tau_o$	5×10^{-2}	Irrelevant
Background degeneracy	2×10^{-8}	Irrelevant
Predetection filter bandwidth	0.5 Å 6×10^{10} Hz)	
Gain bandwidth of amplifier		5×10^7 Hz
Receiver field of view (atmosphere limited)	$\Omega = 10^{-8}$ sterad	
Signal coherence diameter		1 m
Atmospheric transmissivity	0.7	0.7

9.3.2 Satellite Receiver

For a satellite-borne receiver, the analysis carried out previously suggests a comparison of a heterodyne receiver (H) at 10μ with a direct-detection receiver (D) at $0.5\,\mu$. The ratio of required transmitter powers becomes

$$\frac{P_t(0.5\mu)}{P_t(10\mu)} = \frac{(\eta\tau_o)_H A_{RH} \lambda_D}{(\eta\tau_o)_D A_{RD} \lambda_H} \left[\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{(\eta\tau_o)_D \delta NW_D}{5H}}\right] \qquad (70)$$

In this equation it has been assumed that (since there is no atmospheric transmission) the coherence area at $10\,\mu$ is at least as large as the receiver aperture employed. For a satellite receiver N, the number of spatial modes in the background radiation, will be determined, as discussed in Section 1, by the particulars of the mission. It is given by

$$N = \frac{A_{RD}}{\lambda^2} \Omega$$

where Ω is the solid angle subtended by the background radiation. Assume $H = 10^6$, $A_{RD} = 10$ square meters, and take $\delta = 1.3 \times 10^{-8}$ (Section 1) and the other parameters as before. Then one can define Ω_{eff} by the equation

$$\frac{(\eta\tau_o)_D \delta N W_D}{5H} \equiv \frac{\Omega}{\Omega_{eff}}$$

and calculate $\Omega_{eff} = 3 \times 10^{-9}$ sterad. Thus, if $\Omega \lesssim 10^{-9}$ sterad, one can neglect Ω/Ω_{eff}. That is, if the background radiation source does not subtend a greater solid angle than this, then the direct-detection receiver is virtually in the shot-noise limit. Coincidentally, Ω_{eff} is approximately the solid angle subtended by Mars at a distance of 10^8 miles from the earth. Therefore, a small error will usually be incurred if one neglects the background radiation in Equation (70), leaving

$$\frac{P_t(0.5\mu)}{P_t(10\mu)} \approx \frac{(\eta\tau_o)_H}{(\eta\tau_o)_D} \frac{A_{RH}}{A_{RD}} \frac{\lambda_D}{\lambda_H}$$

One might reasonably assume that the receiver area used is the same in both detectors, since the size will be limited by difficulties of a mechanical nature. Thus

$$\frac{P_t(0.5\mu)}{P_t(10\mu)} \approx \frac{1}{20} \frac{(\eta\tau_o)_H}{(\eta\tau_o)_D}$$

Using as estimates $(\eta\tau_o)_D = 5 \times 10^{-2}$, $(\eta\tau_o)_H = 0.2$ gives the result that the 10μ transmitter requires 5 times as much power as the 0.5μ transmitter to achieve $H = 10^6$ bits per second information rate to a satellite receiver. This is about the same ratio that was found for a ground receiver. Therefore, as was pointed out before, the 10μ system probably requires somewhat less power in the space vehicle because of larger power efficiency. The absolute values of transmitter power are listed in the following table:*

Range (mi)	$P_t\ (0.5\mu)$ (watts)	$P_t\ (10\mu)$ (watts)
10^8	0.14	0.5
10^9	10	50

The assumptions involved are listed below.

Assumptions

	0.5μ	10μ
Transmitter diameter	1 m	1 m
Receiver area	10 m^2	10 m^2
Detector efficiency, η_{τ_o}	5×10^{-2}	0.2
Mars background degeneracy	1.3^{-8}	–

9.4 Discussion

In the above calculations many assumptions of a qualitative nature, and many quantitative estimates, have been made. First, calculations were made on the basis of specific binary modulation schemes. The relative merits of the different systems should be nearly insensitive to the choice of modulation, but the absolute values of transmitter powers required are, of course, dependent on the choice. Then choices of system parameters were made, such as receiver efficiencies and aperture sizes. Here the comparisons do depend on the choices of these parameters but it is not expected that they are extremely sensitive to changes which are not large, say, within an order of magnitude. On the other hand, $H = 10^6$ bits per second has been used throughout, and the systems comparisons are quite sensitive to the desired information rate. For example, as H gets very small, the signal powers required become small, the effect of background radiation becomes large and the heterodyne system tends to be favored. If H becomes vary large, the signal power required becomes large and finally the preamplifier receiver tends to become favored. Thus, all the results contained here must be reconsidered in the light of changing communication requirements.

As has been emphasized, the foregoing considerations have been for some quite idealized circumstances. It has been assumed that it is possible to design an efficient 10μ amplifier with high gain and the desired directional sensitivity, to build large aperture receivers and maintain them with necessary tolerances, to compensate for Doppler shifts and operate a large aperture heterodyne system, to point a narrow beam with great accuracy, etc. These assumptions are evaluated in other sections of this report, principally in Chapter 3. The foregoing analysis, in fact, tends to show that while optical communications at megabit rates from

*In calculating this table, it was not assumed that the direct-detection receiver is shot-noise limited at a range of 10^8 miles. This accounts for the deviation from the factor of 5 computed above for the ratio of transmitter powers.

distances of 1 to 10 AU is feasib.. .rom a communication performance point of view, the choice of system is probably not clearly determined from a comparison of idealized systems performance a one.

REFERENCES

1. Investigation of Optical Spectral Regions for Space Communications, ASD Technical Documentary Report No. 63-185, A.F. Avionics Laboratory, Wright-Patterson Air Force Base (May, 1963).
2. A.T. Forrester, R.A. Gudmundsen, and P.O. Johnson, "Photoelectric Mixing of Incoherent Light," Phys. Rev., Vol. 99, No. 6 (September 15, 1955), p 1691.
3. R. Hanbury Brown and R.Q. Twiss. "Interferometry of the Intensity Fluctuations of Light," Proc. Roy. Soc. A, 242 (1957), p 300.
4. Determination of Optical Technology Experiments for a Satellite, Phase I Report, Engineering Report 7846, Perkin-Elmer Electro-Optical Division (July-November 1964).
5. L. Mandel, "Fluctuations of Photon Beams: The Distribution of the Photoelectrons." Proc. Phys. Soc., 74 (1959), p 233.
6. J.I. Bowen, "On the Capacity of a Noiseless Photon Channel," IEEE Trans. Inf. Theory, IT-13 No. 2 (1967), p 230.
7. E.M. Purcell, Nature, 178 (1956), p 1449.
8. A.A.M. Saleh, "An Investigation of Laser Wave Depolarization by Atmospheric Transmission," IEEE Conf. on Laser Eng. and Appl., Conf. Digest (June 1967), p 32.
9. A.L. Buck, "Effects of the Atmosphere on Laser Beam Propagation," Appl. Optics, Vol. 6, (April 1967), p 703.
10. T.S. Chu, "Attenuation by Precipitation of Laser Beams at 0.63μ, 3.5μ, and 10.6μ ," IEEE Conf. on Laser Eng. and Appl., Conf. Digest (June 1967), p 30.
11. D.J. Blattner and L. Bordogna, RCA Report AD276526, "Research Program on the Utilization of Coherent Light" (January 1 – March 31, 1962).
12. I.P. Peteranecz and R.A. Simons, "Atmospheric Propagation Studies at Optical Millimeter, and Microwave Frequencies - Part I," Air Force Avionics Laboratory Report AFAL - TR-65-79 (March 1965).
13. P.B. Taylor, Atmospheric Propagation Studies at Optical, Millimeter, and Microwave Frequencies - Part II, Air Force Avionics Laboratory Report AFAL-TR-65-79 (March 1965).
14. T.S. Chu, "Measurements of Optical Beams Propagated through the Atmosphere," 1965 IEEE Antennas and Propagation Symposium Digest, p 117.
15. R.F. Lucy, K. Lang, C.J. Peters, and K. Duval, "Optical Superheterodyne Receiver," Appl. Optics, 6 (August 1967), p 1333.
16. M. Subramanian and J.A. Collinson, "Modulation of Laser Beams by Atmospheric Turbulence - Depth of Modulation," BSTJ., XLVI (March 1967), p 623.
17. D.L. Fried, G.E. Mevers, and M.P. Keister, Jr., "Measurements of Laser Beam Scintillation in the Atmosphere," JOSA, 57 (June 1967), p 787.
18. Private communication with R.F. Lucy.
19. Private communication with R.D. Rosner, Bell Telephone Laboratories.
20. F.E. Goodwin, "A 3.39 - Micron Infrared Optical Heterodyne Communication System," IEEE Conf on Laser Eng. and Appl., Conf. Digest (June 1967), p 28.
21. D. H. Höhn, "Effects of Atmospheric Turbulence on the Transmission of a Laser Beam at 6328Å. I – Distribution of Intensity," Appl. Opt. 5 (September 1966), p 1427.
22. D. H. Höhn, "Effects of Atmospheric Turbulence on the Transmission of a Laser Beam at 6328Å. II – Frequency Spectra," Appl. Opt., 5 (September 1966), p 1433.
23. T. S. Chu, "On the Wavelength Dependence of the Spectrum of Laser Beams Traversing the Atmosphere," Appl. Optics 6 (January 1967), p 163.
24. D. C. Hogg, "On the Spectrum of O,tical Waves Propagated through the Atmosphere," BSTJ, XLII (November 1963), p 2967.
25. M. Subramanian and J. A. Collinson, "Modulation of Laser Beams by Atmospheric Turbulence," BSTJ, XLIV (March 1965), p 543.
26. V. I. Tatarski, Wave Propagation in a Turbulent Medium, R. A. Silverman, translator (New York, McGraw-Hill, 1961).
27. I. Goldstein, A. Chabot, and P. A. Miles, "Heterodyne Measurements of Light Propagation Through Atmospheric Turbulence," Proc. IEEE, 53 (September 1965), p 1172.
28. W. R. Hinchman and A. L. Buck, "Fluctuations in a Laser Beam over 9- and 90-Mile Paths," Proc. IEEE, 52 (March 1964) p 305.
29. S. H. Reiger, "Starlight Scintillation and Atmospheric Turbulence," Astron, Jour., 68 (August 1963), p 395.
30. J. B. Keller, "Stochastic Equations and Wave Propagation in Random Media," Proc. Symp. in Appl. Math., XVI (1964), p 145.
31. L. S. Taylor, "Decay of Mutual Coherence in Turbulent Media," JOSA, 57 (March 1967), p 304.

32. H. C. Van de Hulst, Light Scattering by Small Particles, Ch. 11 (New York, John Wiley, 1957).

33. P. Beckmann, "Signal Degeneration in Laser Beams Propagated Through a Turbulent Atmosphere," Radio Sci., 69D (April 1965), p 629.

34. L. Elterman, "Parameters for Attenuation in the Atmospheric Windows for Fifteen Wavelengths," Applied Opt, 3 (June 1964), p 745.

35. R. E. Hufnagel and N. R. Stanley, "Modulation Transfer Function Associated with Image Transmission through Turbulent Media," JOSA, 54 (January 1964), p 52.

36. M. J. Beran, "Propagation of the Mutual Coherence Function through Random Media," JOSA, 56 (November 1966), p 1475.

37. W. P. Brown, "Propagation in Random Media – Cumulative Effect of Weak Inhomogeneities," IEEE Trans, AP-15 (January 1967), p 81.

38. Private communication with R. E. Hufnagel.

39. D. L. Fried, "Optical Heterodyne Detection of an Atmospherically Distorted Signal Wavefront," Proc. IEEE, 55 (January 1967), p 57.

40. D. L. Fried, "Test of the Rytov Approximation," JOSA, 57 (February 1967), p 268.

41. L. A. Chernov, Wave Propagation in a Random Medium, R. A. Silverman, translator (New York, McGraw-Hill, 1960).

42. W. N. Peters and R. J. Arguello, "Fading and Polarization Noise of a PCM/PL System," IEEE Conf. on Laser Eng. and Appl., Conf. Digest (June 1967), p 29.

43. D. L. Fried, "Atmospheric Modulation Noise in an Optical Heterodyne Receiver," IEEE Jour. of Quantum Elec., QE-3 (June 1967), p 213.

44. G. R. Heidbreder, "Image Degradation with Random Wavefront, Tilt Compensation," IEEE Trans., AP-15 (January 1967), p 90.

45. Wood's Hole Summer Study on Restoration of Atmospherically Degraded Images, Vol. 2 (July 1966).

46. D. C. Fosth and W. H. Zimmerman, "High Accuracy Attitude Control for Space Astronomy," JACC (1967), Preprint.

47. M. S. Lipsett, Laser/Optics Techniques, 2nd Interim Summary Report, Engineering Report No. 8631, Perkin-Elmer Optical Group, Norwalk, Connecticut, prepared under NASA Contract No. NAS8-20115.

48. A. Wallace, Study of Laser Pointing Problems, Kollsman Instrument Corp. Report KIC-RD-000162-2, NASA CR-60699, prepared under NASA Contract No. NASW-929.

49. C. S. Proste, Echelon Control System, Ph. D. Thesis, Texas A & M (1967).

50. B. T. Bachafer and L. T. Seaman, "One-Arc-Second Simulator for Orbiting Astronomical Observatory," Journal of Spacecraft and Rockets, Vol. 2, No. 2, pp 260-262.

51. A. E. Lopez et al, "Results of Studies on a Twin-Gyro Attitude-Control System for Space Vehicles," Journal of Spacecraft and Rockets, Vol. 1, No. 4, pp 399-402.

52. G. C. Newton et al, Analytical Design of Linear Feedback Controls (New York, John Wiley, 1957) p 372.

53. J.T. Tou, Digital and Sampled Data Control Systems, (New York, McGraw-Hill, 1959), pp 349-356.

54. G. Biernson and R. F. Lucy, "Requirements of a Coherent Laser Pulse-Doppler Radar," Proc. IEEE, 51 (January 1963), p 202.

55. D. L. Fried and R. A. Schmeltzer, "The Effect of Atmospheric Scintillation on an Optical Data Channel-Laser Radar and Binary Communication," Appl. Opt., 6, No. 10 (October 1967), p 1724.

56. A. E. Siegman, "The Antenna Properties of Optical Heterodyne Receivers," Proc. IEEE, 54 (October 1966), p 1350.

57. Monte Ross, "Pulse Interval Modulation (PIM) Laser Communications," Eastcon Tech. Convention Record (1967).

58. B. M. Oliver, "Signal-to-Noise Ratios in Photoelectric Mixing," Proc. IRE, 49 (December 1961), p 1960.

59. H. A. Haus and C. H. Townes, "Comments on Noise in Photoelectric Mixing," Proc. IRE, 50 (June 1962), p 1544.

60. J. Rodda and L. H. Enloe, Proc. IEEE, 53 (2) (1965), pp 165-166.

61. B. Cooper, "Optical Communications in the Earth's Atmosphere," IEEE Spectrum (July 1966).

62. Private communication with D. H. Ring, Bell Telephone Laboratories.

63. L. K. Anderson and B. J. McMurtry, "High Speed Photodetectors," Proc. IEEE, 54, No. 10 (1966), p 1335.

64. D. E. McCumber, "Intensity Fluctuations in the Output of CW Laser Oscillators I," Phys. Rev., 141 (1966), p 306.

65. H. Kogelnik and A. Yariv, "Considerations of Noise and Schemes for its Reduction in Laser Amplifiers," Proc. IEEE, 52, No. 2 (February 1964), pp 165-172.

66. B. M. Oliver, "Thermal and Quantum Noise," Proc. IEEE, 53, No. 5 (May 1965), pp 436-454.

67. R. W. Koepcke, "On the Control of Linear Systems with Pure Time Delay," ASME Journ. Basic Engr. (March 1965), p 74.

68. J. P. Gordon, "Quantum Effects in Communication Systems," Proc. IRE, 50 (September 1962), pp 1896-1908.

New Technology

PULSE POSITION MODULATION FOR OPTICAL COMMUNICATION

In optical communication systems there are three fundamental sources of noise which can limit the communication capacity. The most basic is the shot noise in the detected signal current which is sometimes called quantum noise. Except for rather impractical conditions of very wide bands, very small signals, and the absence of all other noise, this may be treated as white Gaussian noise with a signal-to-noise ratio of $P_s/2hv$ per cycle, which for $\lambda = 1\mu$, is 187 dB per watt. The second noise source is the FKT noise of the amplifier following the optical detector. Its importance depends on the detected signal level delivered by the photo-detector. It can be overcome by noiseless gain in the photo-detector and, since photodetectors are square-law devices, the signal-to-FKT-noise ratio is proportional to the square of the optical signal power. Noiseless gain can be realized by heterodyne detection, photomultipliers, and to some extent by solid state avalanche detectors. The third source of noise in many communication applications is background optical radiation reaching the detector with the signal. The background radiation is detected like the signal and produces output current with its accompanying shot noise as well as beat product noise components due to self beats and beats with the signal. Only the beat with the signal is a function of signal level and it, as well as the background self beat noise, is usually small relative to the shot noise. Thus, a given background produces a fixed noise level independent of the signal.

In optical systems the significance of output signal-to-noise ratio as a figure of merit can be different from the classical signal-to- constant-added-noise ratio. This happens when both the FKT noise and the background noise, which are constant, are dominated by or comparable to the signal shot noise. Thus, with on-off pulse signal modulation with zero background and zero FKT noise, a given signal-to-noise ratio on the received pulses would produce the same "on" errors as a 6 dB better signal-to-noise ratio when the same noise is present during both the on and off periods, and would produce no "off" errors at all. When background noise dominates in an optical power detector system, the output signal-to-noise ratio is not proportional to received signal power but to its square. Thus a reduction in the required signal-to-noise ratio because of reduced bandwidth, error correction, etc., by 1 dB results in only 1/2 dB reduction in the required transmitted power.

There have been a number of proposals to use pulse position modulation (PPM) for optical communication[1]. The analysis has been developed from the basic Poisson distribution of photoelectrons with the implication that the gain achieved is due to this "quantum" treatment of communication signals. A more practical view would seem to be that the quantum gain and, in many cases, even the gain due to the absence of signal shot noise in off periods (or its variation with signal intensity) are small or non-existent because of the presence of constant system noise which is independent of signal level. This is particularly true if photomultipliers cannot be used. Heterodyne detection does not have these gains because the noise in the output is constant regardless of signal, although a signal-to-noise ratio of P_s/hv is ideally possible. In spite of the above remarks, PPM can yield gains in effective signal-to-noise ratio or reduced error rates for the same average signal power in the absence of any quantum effects. This gain results from the square-law detection, the use of a much wider system bandwidth, and the PPM. It is effective against constant added noise from any source.

An efficient simple optical communication system would utilize binary pulse code modulation (PCM) with switching between the two polarizations at the transmitter, a receiving channel for each polarization at the receiver, and a differential decision device yielding a plus-minus binary baseband output. The error rate vs. signal and background noise levels for such a system has been calculated in Reference 2 using the Poisson distribution of received light photons to determine the signal output fluctuations or noise. The calculation has also been made with the assumption that the noise is represented by "white" noise with a Gaussian distribution and the same power. With the pure Poisson distribution when the total receiver output contains no background noise of any kind, a given small ($<10^{-3}$) error rate requires about 3 dB less signal power than would be calculated using the Gaussian approximation. This difference disappears rapidly in the presence of background noise, however, and for equal signal and background noise powers the difference is less than 1 dB. In almost all practical situations, therefore, the simpler Gaussian noise calculation is quite adequate, and if there is an error it is conservative rather than over-optimistic.

This chapter describes a system which incorporates Q switching, cavity dumping, multiple lasers, PPM, and polarization modulation with the object of increasing the communication capacity of the simple reference system described above in bits per watt of average radiated light power. The system is based on a proposed modulation scheme which should be capable of delivering short PPM pulses with the same average power as the CW pumping system will yield for CW operation for a Nd-Yag laser.

The modulation arrangement will be described and then the gain relative to the simple system due to the increased peak power and square-law detector will be described. This is followed by a discussion of the further gains achieved by PPM combined with polarization modulation and the use of multiple lasers. An optimum amount of PPM is found for maximum system gain. An estimate is then made of the equivalent transmitter power gain of the proposed system over the simple polarization modulated system for the same error rate. Since much of this system gain depends on using a much wider bandwidth for the proposed system, the advantage, in the face of equipment limitations, will decrease for high capacity systems. However, at a bit rate of 10^6, substantial gain should be realizable.

1. PPM LASER MODULATION

Basic to the systems to be discussed is a transmitter capable of delivering the same average power with the same efficiency while the pulse length and duty cycle are reduced and also capable of being pulse position modulated. Such a laser transmitter has not been described as far as is known, but a combination of known techniques to produce the desired characteristics appears to be possible. Combined Q switching and cavity dumping have been used to achieve high peak powers.[3,4] In the arrangements described, flash lamp pumps and separate controls for the Q switch and dumping functions were used. Such arrangements are not suitable for high speed pulse position modulation.

However, if we start with the Nd-Yag laser which is capable of CW operation, and provide a cavity dumping polarization modulator calcite-prism combination as shown in Figure A, the desired operation should result from simply reversing the usual practice of pulsing the dumping modulator to produce an output pulse. If the dumping circuit is biased to enable the dump path for no applied modulation voltage, then a short pulse of suitable length will produce the effects indicated in Figure B. Before the signal pulse is applied, the laser rod is coupled to the output and has no cavity. During this period the CW pump is pumping atoms in the active material or "charging" the laser. In the time scale of interest and the absence of stimulation by cavity radiation, most of the atoms inverted will remain so until the signal pulse switches the cavity "on." Laser oscillation then starts to build up and reaches a peak at some time characteristic of the system Q, pump power, and charging time. The length of the modulation pulse is adjusted so that the modulating pulse falls to zero at the peak of the laser buildup.[3] When it goes to zero it re-establishes the output path and the peak laser energy is dumped as an output pulse. The laser then remains inactive, but charging, until the next modulation pulse comes along.

The length and amplitude of the output pulse will depend on the fall time of the modulation pulse and the cavity length. Zero fall time would yield a pulse length of 2L/C if L is the cavity length and C the speed of light. The output pulse is lengthened as the modulation fall time becomes finite and can be adjusted over a limited range in this way. The laser buildup time will depend on the pump power, cavity losses, and cavity length. The arrangement appears to be sufficiently flexible so that the CW average power capability of a pump-laser rod combination can be approached with short pulses at relatively high repetition rates by optimizing the cavity length and modulating pulse shape and duration. In the example of the proposed system described in this chapter, we are interested in a pulse length of $5x10^{-9}$ seconds, and a repetition rate of about $5x10^4$. The charge time would vary from 17 to $23x10^{-6}$ seconds.

2. SQUARE-LAW GAIN

Consider a pulse modulated communication system with a capacity of H bits per second. Mention of "gain" throughout the discussion will usually refer to gain with respect to a reference system in which the average power output, P_o, of the transmitter laser is polarization modulated by a binary signal. After attenuation by transmission, the average signal power, P_n, arrives at the receiver detector. In the reference system the average power and the pulse peak power are the same. In the other systems discussed, the received pulse peak power is P_m, but the average power is equal to the P_n of the reference system.

The transmitter source is assumed to have the property that the average power delivered, P_o, remains constant and independent of the pulse duration when it is pulse modulated. The Q-switched Nd-Yag laser with a constant

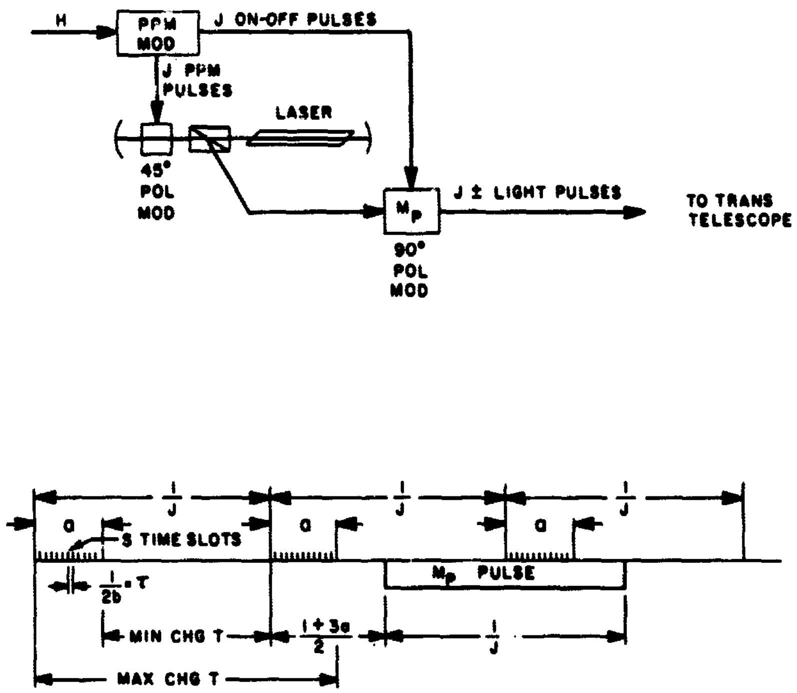


Figure A. Single laser combined PPM and polarization modulated CW pumped transmitter

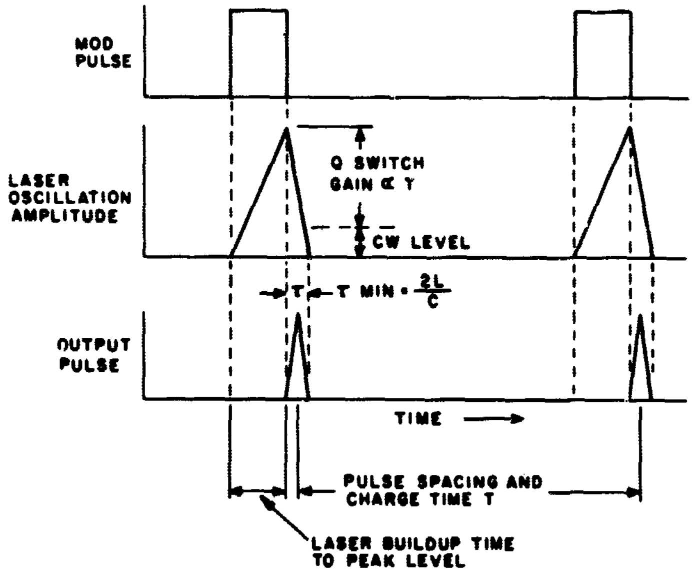

MOD PULSE
LASER OSCILLATION AMPLITUDE
Q SWITCH GAIN ∝ T
CW LEVEL
T
T MIN = $\frac{2L}{C}$
OUTPUT PULSE
TIME
PULSE SPACING AND CHARGE TIME T
LASER BUILDUP TIME TO PEAK LEVEL

Figure B. Qualitative timing sequence for combined Q switching – cavity dumping a CW pumped laser pump source that has been described should approximate this performance. If the pulses come at a constant rate, J per second, the received peak pulse power, P_m, will be inversely proportional to J and the pulse length τ. Thus,

$$P_m = \frac{P_n}{J\tau} = \frac{2bP_n}{J} \qquad (1)$$

where $b = 1/2\tau$ is the minimum bandwidth required to receive the pulses without loss.

In the output of a photo-detector normalized to a 1 ohm impedance level, the peak pulse power, S, will be

$$S = \left(\frac{\eta e P_m}{hv}\right)^2 = \left(\frac{2\eta e P_n b}{hvJ}\right)^2 \qquad (2)$$

where η is the detector quantum efficiency, e is the electronic charge, h is Planck's constant, and v is the optical frequency. The signal shot noise power when the pulse is present will be

$$N_{sh} = \frac{4\eta e^2 P_n b^2}{hvJ}$$

and the background noise power will be

$$N_B = N_o b$$

where N_o is the total background noise density resulting from received background radiation shot noise, dark current shot noise, and FKT noise. Then the signal-to-signal-shot-noise ratio is

$$\frac{S}{N_{sh}} = \frac{\eta P_n}{hvJ} \qquad (3)$$

and the signal-to-background noise is

$$\frac{S}{N_B} = \left(\frac{2\eta e P_n}{hvJ}\right)^2 \frac{b}{N_o} = \left(\frac{2\eta e P_n}{hvJ}\right)^2 \frac{1}{2N_o\tau} \qquad (4)$$

Note that when the bandwidth is adjusted to fit the pulse width and the average power is independent of the pulse width, the signal-to-signal-shot-noise ratio is independent of the pulse width or bandwidth. However, the signal-to-background-noise ratio is directly proportional to bandwidth or inversely proportional to the pulse width. Thus the tolerance of pulse systems in general for all types of constant background noise will be improved under the assumed conditions by maximizing the bandwidth, and this improvement will not entail any penalty with respect to signal shot noise. This improvement is referred to as the square-law gain.

3. PPM GAIN

If b is made as large as possible, further gain can be achieved by reducing the pulse rate J in Equations (3) and (4) to a smaller value if this is done without reducing the information capacity. This can be accomplished by utilizing pulse position modulation (PPM) to derive more than one bit of information from each pulse. If a synchronous system is assumed, then each pulse can be made to occur in any one of S time positions. There are $1/J\tau$ resolvable time slots available between the unmodulated pulses. If we utilize S of these possible time slots then each of the J pulses can represent R bits where

$$R = \text{Log}_2 S \qquad (5)$$

and

$$J = H/R \qquad (6)$$

The smallest possible J will result if all the resolvable time slots are utilized. However, if this is done, then there is a possible modulation sequence calling for two of the J pulses to occur in adjacent time slots. This would allow no charge time for the Q switch laser and no second pulse would occur. Therefore it is necessary to limit the portion of each 1/J time period in which pulses are permitted to allow some minimum charge time for the laser. This unused time will determine the amplitude of the minimum transmitted pulse. There will be some optimum J that will deliver the maximum minimum pulse power for a system with capacity H and a given bandwidth that determines the resolving time. For this optimum pulse rate J, only a portion a of the pulse period will be required for pulses and the remainder will be available for charging the laser. This is illustrated by the timing diagram in Figure A.

There are other modifications that are available for improving the system performance. Since charge time determines pulse amplitude, a simple way to increase this would be to use multiple lasers with successive pulses coming from each of L lasers in sequence. Each laser would then be responsible for H/L bits of system capacity, the average transmitted power would be increased by L times, and the minimum peak power would be somewhat more than L times greater. Conversely, the unused time reserved for charging a single laser system can be filled in with pulse trains from multiple lasers to increase the system bit capacity at the same performance level without the need for additional transmission paths.

These advantages can readily be obtained passively by combining two lasers without loss by utilizing the two orthogonal polarizations and calcite prisms. If more than two lasers of the same frequency are to be combined into a single optical system without loss, time multiplexing with active switches is quite feasible with polarization modulators. In a tree arrangement starting with calcite prisms for each pair of lasers, two polarization modulators could combine four sources, four could combine eight, and so on. These polarization combining schemes will yield an output signal on both plus and minus polarizations. Another polarization modulator can be added to the combined pulse stream to yield all plus or all minus pulses for a single polarization signal and thus save one receiver channel.

Since two transmission polarizations are available, they can be utilized for further improvement of the system. If the laser is followed by a polarization modulator, then each pulse can be either plus or minus polarization as demanded by the encoding logic. If each pulse has two possible values, then only half as many pulse positions, S, are required. Reducing the required number of used time slots will reduce the active time and increase the charge time and pulse amplitude. It will also increase the number of lasers that can be time multiplexed. Each J pulse represents R bits. The use of two polarizations in this way adds one bit to R at a cost of doubling the error rate, because any single pulse may appear in either of the two required receiver channels each of which contributes errors. The increase in error rate for the same pulse amplitude will be more than compensated for by the reduction in J and the resulting increased pulse amplitude. The net gain will be greater for small R.

4. COMPLETE SYSTEMS

Figure A shows both a functional diagram indicating the major components for a 1 laser PPM system and also a timing diagram for the resulting pulse train. The polarization modulator in the output light path may be omitted if the gain from using two polarizations is not considered worthwhile.

Figure C is a functional diagram of a receiver suitable for this PPM system. If only one polarization is used, one of the identical parallel paths is removed. It may be that more sophisticated and effective noise reducing circuits for quantized PPM are available, but the processing indicated in Figure C should yield reasonably good results. The incoming pulses are first sliced so that only pulses above a threshold level pass the slicer. Beyond the slicer a time gate is indicated which passes only pulses occurring during the active part $\underline{a}$, of the cycle. This eliminates any false pulses coming when none could be transmitted.

The remaining pulses go to an error reducing circuit which examines the interval $\underline{a}$ to determine if more than one pulse has been received. If so, it selects the highest received pulse as the signal for that interval. At this point further information can be obtained if the time between the present pulses and the last preceding pulse is measured. Since the amplitudes of the transmitted pulses will vary somewhat depending on their charging time, this time can be used to weight the received multiple pulses in making a choice. Having selected one pulse per interval, its position is interpreted as one of S numbers from which H/J of the original information bits can be derived.

All the above operations must be carried out with pulses of τ width, or bandwidth b, and the speed at which the necessary logic can operate may well be the limiting factor in the system, since there is gain from greater b.

In this system low error rates can be achieved with the average background noise power received, P_B, much greater than the average signal power, P_n. This means that the dc photocurrent will be proportional to P_B. The maximum range for a given transmitter power and error rate will be achieved with a precise setting of the slicer level. This optimum level depends on the prevailing background noise level which may change with time. Thus the slicer level, or the linear signal gain preceding the slicer, must be controlled by the dc photocurrent.

Figure D shows a two laser transmitter arrangement which gives somewhat more gain than that caused by simply doubling the average power of a single laser system. The combination of the two laser outputs results in a two polarization signal with alternate plus and minus pulses. Either the polarization modulator shown can be used to invert one set of pulses to reduce the system to a one polarization system with a smaller error rate, or the plus-minus output can be retained and the polarization of each transmitted pulse controlled by the encoding logic. The latter arrangement will improve performance somewhat at the cost of more complex encoding-decoding. This transmitter requires the same receiver system as the single laser transmitter.

5. OPTIMIZING THE PPM DESIGN

Assume that a capacity of H binary bits is required and that the maximum bandwidth that can be used is b Hz. The objective is to find the values of the number of time slots S, the pulse repetition rate per laser J, and the proportion of time $\underline{a}$ allotted to the PPM, as shown in Figures A and D, that will maximize the minimum transmitted pulse peak power P for a system using L lasers with an average transmitted power of P_o watts per laser.

In a multiple laser arrangement the bits per laser will be H/L. If each transmitted pulse has a value of R bits, then the number of pulses transmitted per laser will be

$$RJ = \frac{H}{L} \text{ and } R = \frac{H}{LJ} \tag{7}$$

The number of time slots required for a pulse to convey R bits is

$$S = 2^R = 2^{H/LJ} \tag{8}$$

In Figures A and D means are provided for sending each pulse as either plus or minus polarization. Thus each pulse position can actually represent two states. Other means, such as amplitude modulation or different frequencies, might be used to increase the number of states represented by a single pulse in a time slot. The number of

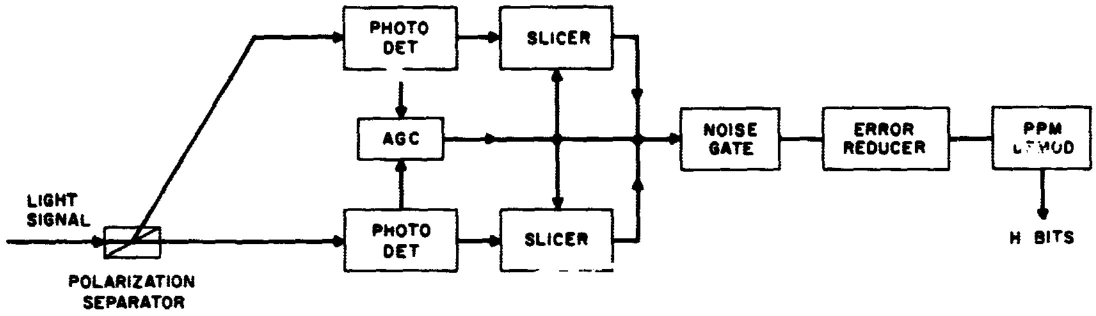


Figure C. Receiver for combined PPM and polarization modulation system

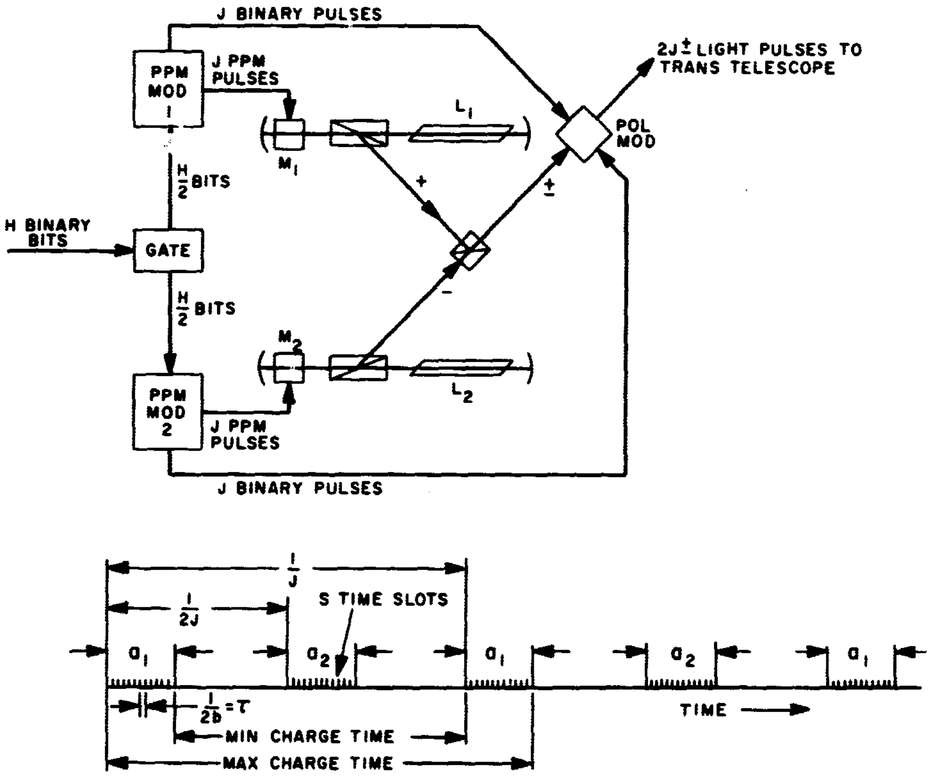


Figure D. Two laser combined system

time slots required is reduced by the number of states represented by each pulse. Using β for the number of states per pulse, Equation (8) becomes

$$S = \frac{2^R}{\beta} = \frac{2^{H/LJ}}{\beta} \tag{9}$$

and

$$R = \log_2 S\beta \tag{10}$$

From Equation (7)

$$J = \frac{H}{L \log_2 S\beta} \tag{11}$$

The width of the individual time slots is $\tau = 1/2b$ seconds. The total time occupied by the S slots is

$$S\tau = \frac{S}{2b} \tag{12}$$

Since the pulse period is 1/J, the portion of the period occupied by the S positions is

$$a = \frac{JS}{2b} = \frac{SH}{2bL \log_2 S\beta} \tag{13}$$

For constant average power from the laser, the peak pulse power is given by

$$P = \frac{P_o}{J\tau} = \frac{P_o 2b}{J}$$

if the charge time is I/J. Figure D shows that, when PPM is used, the minimum charge time is (1-a)/J and the maximum is (1+a)/J. Then the peak power of the pulses will be

$$P_{av} = \frac{P_o 2b}{J}, P_{min} = \frac{P_o 2b(1-a)}{J}, P_{max} = \frac{P_o 2b(1+a)}{J} \tag{14}$$

and

$$\frac{P_{min}}{P_{av}} = (1-a) \qquad \frac{P_{max}}{P_{av}} = (1+a) \tag{15}$$

An optimum system design will be one in which the minimum pulse is maximized. Using Equations (9), (13), and (14)

$$\frac{P_{min}}{P_o} = \left(\frac{2b}{J} - \frac{2^{H/LJ}}{\beta}\right) \tag{16}$$

Setting the derivative of P_{min}/P_o with respect to J equal to zero yields

$$J_m = \frac{H}{L \log_2 \left(\frac{2bL\beta}{H \log_e 2}\right)} = \frac{H}{L \log_2 \left(\frac{2.89bL\beta}{H}\right)} \tag{17}$$

Substituting Equation (17) in Equations (16), (13), (9), and (10) yields

$$\frac{P_{min}}{P_o} = \frac{2bL}{H}\left(\log_2 \frac{2.89bL\beta}{H} - 1.445\right) \tag{18}$$

$$a = \frac{1.445}{\log_2 \frac{2.89Lb\beta}{H}} \tag{19}$$

$$S = \frac{2.89Lb}{H} \tag{20}$$

$$R = \log_2 \frac{2.89Lb\beta}{H} \tag{21}$$

Equations (17) to (21) give the parameters of a modified PPM system optimized to yield the highest amplitude minimum pulse. Table A lists the values of the various parameters calculated for systems utilizing different numbers of lasers with and without polarization modulation in some cases. For all systems listed in Table A the receiver bandwidth is assumed to be 10^8 Hz, and the information rate 10^6 bits per second. The 16 laser system is not possible because aL>1, but more than 8 lasers could be used without reducing a below the optimum value. It will be noted that the peak pulse power increases somewhat more than linearly as more lasers are used. In a linear system, this would presumably be compensated for by the greater signal-to-noise ratio required by the increase in S with the number of lasers, but in the square-law optical system the increase in pulse power should yield a real performance gain. The system parameters were calculated to yield a maximum minimum pulse, but the ratio of the average modulated pulse to the laser average power for each system is listed in Table A.

Table A

MULTI-LASER SYSTEM PARAMETERS

L	β	R	$JX10^{-4}$	S	a	$\frac{1+a}{1-a}$	$\frac{P_{av}}{P_o}$
1	1	8.17	12.22	289	0.177	1.43	1,640
1	2	9.17	10.90	289	0.157	1.38	1,833
2	1	9.17	5.44	578	0.157	1.38	3,670
2	2	10.17	4.91	578	0.142	1.33	4,070
4	2	11.17	2.24	1,156	0.129	1.30	8,930
8	2	12.17	1.03	2,312	0.119	1.27	19,400
16	2	13.17	0.474	4,624	0.110	1.25	42,200

Information rate, H = 10^6 bits/second
Receiver bandwidth, $b_m = 10^8$ Hz
L = Number of lasers in transmitter
β = Number of pulse states (polarization)
R = Bits per transmitted pulse
J = Pulse rate per laser
S = PPM time slots used per laser
a = Proportion of laser pulse period used for PPM
P_{av} = Peak power of average modulated pulse
P_o = Average power per laser

6. POWER GAIN OF PROPOSED SYSTEM

An exact analysis of the error rate performance of the proposed system is far beyond the scope of this chapter, but it is possible to estimate the effective transmitter power gain of the system to within 1 or 2 dB quite simply with respect to the well-documented system of Reference 1. This can be done by assuming Gaussian noise and depending upon the fact that for useful error rates the error rate varies more than an order of magnitude per dB change in signal-to-noise ratio for all kinds of noise of interest.

The Gaussian noise approximation calls for about 3 dB more signal-to-noise ratio than the true Poisson noise in the shot noise limited case with no background noise. As background noise is added, this difference drops to less than 1 dB when the shot noise and background are equal, and is practically negligible for greater background. Background noise is understood to mean all constant noise including background radiation, FKT noise, and dark current noise. The steep error curve will be invoked in estimating the performance of the proposed system.

The reference system is a binary polarization modulated system which delivers a peak pulse power P_n to the detection system along with a total background optical power of P_B. The receiver separates the polarizations and feeds each polarization component to a separate photo-detector. The outputs of the detectors are operated differentially so that in each time slot there is signal shot noise plus background shot noise from both polarizations (full P_B shot noise) and either a positive or negative signal. The bit capacity is H and the bandwidth $b_n = H/2$.

The system of Figure D has the same bit capacity divided between two lasers each with the same average power P_o as the reference system transmitter. Each laser operates at a pulse rate, J, and both PPM and polarization modulation are used. The receiver, shown in Figure C, is assumed synchronized and has two separate channels for the two polarizations. Differential reception cannot be used here and each channel receives the full transmitter output with a peak power of P_m and background noise $P_B/2$. It is assumed that the optimum slicing level which depends on P_B and P_m in the receiver channels is automatically maintained.

In both the reference and modified PPM systems, the pulses are received by a square-law photo-detector which produces a signal pulse peak current

$$I_S = \frac{\eta P_S e}{h\nu} \qquad (22)$$

Where P_S is the pulse peak power and a shot noise rms pulse peak current, assumed Gaussian,

$$I_N = \sqrt{\frac{2P_S e^2 \eta b}{h\nu}} \qquad (23)$$

The output pulse signal-to-shot-noise ratio is

$$U_{Sh} = \frac{\eta P_S}{2h\nu b} \qquad (24)$$

It is interesting to note that, in the Nyquist limit H = 2b, U is the number of photons per pulse for a binary system.

In the presence of uniform background radiation of N watts per square meter per steradian per Angstrom, the received background power is

$$P_B = \frac{NA\Omega B\lambda^2}{C} \qquad (25)$$

where A is the receiver aperture area, Ω the field of view in steradians, B the pre-detection optical bandwidth in Hz, and C the velocity of light. Other background noise will be contributed by post-detector amplifier FKT noise and the detector dark current. The FKT noise power is 4FKTb/Rg where F is the noise figure of the post-detection amplifier when fed by a source resistance Rg driven by a noiseless infinite impedance signal current source. The dark current shot noise is $2eI_D b$.

In the modified PPM system the pulse is shortened and the bandwidth is increased so that the average received power P_n is held constant. Assuming the minimum bandwidth, the pulse length is $\tau = 1/2b_m$ and the peak pulse power is

$$P_m = \frac{P_n}{J\tau} = \frac{2b_m P_n}{J} \qquad (26)$$

The receiver square-law detector input includes the coherent signal and the received background radiation which have different spacial distributions over the detector area. Taking the spatial distributions into account in evaluating the cross product terms, the signal and rms noise powers in the detector output can be evaluated using Rice's[5] relations for the output of a square-law detector. Adding terms for the dark current shot noise and post-detection amplifier FKT noise, the received pulse signal-to-noise ratio is

$$U_m = \frac{\eta P_n}{h\nu J\left[1 + \frac{2\eta\lambda^2 P_B}{h\nu AB\Omega} + \frac{1}{P_m}\left(\frac{\eta\lambda^2 {P_B}^2}{h\nu AB\Omega} + P_B + \frac{2FKTh\nu}{\eta e^2 Rg} + \frac{I_D h\nu}{e\eta}\right)\right]} \qquad (27)$$

In most cases of practical interest the second and third terms representing detector beat products can be dropped in comparison to the first and fourth terms which represent signal shot noise and radiation background shot noise respectively. The last two terms simply add to P_B if present, and it will therefore be assumed that P_B has been adjusted to take these terms into account if they are significant. Then we have

$$U_m = \frac{\eta P_n}{hvJ\left[1 + \frac{P_B}{P_m}\right]} \qquad (28)$$

For the reference system J = H = $2b_n$, so that

$$U_n = \frac{\eta P_n}{2hvb_n\left[1 + \frac{P_B}{P_n}\right]} \qquad (29)$$

The relations in Equations (28) and (29) cannot be used directly to compare the two systems because the reference system uses a differential decision device while the PPM system must use a slicer decision device. In the presence of P_B the slicer device will require a greater U than the differential device for the same error rate. We will assume that an error rate E is required and that the coherent differential binary U_n required by the reference system for this error rate is known. We will estimate the U_m required for a slicer system with the same error rate per pulse time slot.

The comparison is made first for the case in which a slicer is used to make the decision in a simple binary on-off modulation system adjusted for equal mark and space errors compared to a differential system with the same error rate per pulse. The required pulse power will be found in the two cases. The pulse power P_n required in the balanced case, to give the signal-to-noise ratio U_n corresponding to an error rate E, is from Equation (29)

$$P_n = \frac{K_1 U_n}{2}\left[1 + \sqrt{1 + \frac{4P_B}{K_1 U_n}}\right] \qquad (30)$$

$$K_1 = \frac{2hvb_n}{\eta}$$

In the on-off pulse system with a slicer, the slicer level must be adjusted so that the mark and space errors are equal, although the noise in the two intervals is different because there is no signal shot noise in a space interval. When no pulse is present, only background noise appears in the detector output and the slicer current reference W is set to exceed the rms value of the background noise current by the factor $\sqrt{U_n}$ which is the rms signal-to-noise current ratio required for the error rate E. The rms shot noise current due to background radiation is given by Equation (23) if $P_B/2$ is substituted for P_s. Then

$$\sqrt{U_n} = \frac{W}{\sqrt{\frac{P_B e^2 \eta b_m}{hv}}} \qquad W = \sqrt{\frac{K_2 U_n P_B}{2}} \qquad (31)$$

$$K_2 = \frac{2e^2 \eta b_m}{hv}$$

We have used $P_B/2$ rather than P_B because the slicer system requires only one polarization. The bandwidth of the on-off system is b_m and is one-half the pulse rate as before.

For a mark signal, when a pulse with a peak power P_m is present, the slicing level current W is subtracted from the signal current and the available signal current, using Equations (22) and (31), is

$$I_a = (K_3P_m\text{-}W) = K_3P_m - \sqrt{\frac{K_2 U_n P_B}{2}} \qquad (32)$$

$$K_3 = \frac{\eta e}{hv}$$

The rms noise current I_n in a mark interval is the current corresponding to the sum of the signal and background shot noises

$$I_n = \sqrt{K_2\left(P_m + \frac{P_B}{2}\right)}$$

Then

$$\sqrt{U_n} = \frac{I_a}{I_n} = \frac{K_3P_m - \sqrt{\frac{K_2 U_n P_B}{2}}}{\sqrt{K_2\left(P_m + \frac{P_B}{2}\right)}} \qquad (33)$$

Solving Equation (33) for P_m yields

$$P_m = K_4 U_n\left(1 + \sqrt{\frac{2P_B}{K_4 U_n}}\right) \qquad (34)$$

$$K_4 = \frac{K_2}{K_3} = \frac{2hvb_m}{\eta}$$

Then the ratio of the peak pulse power P_m required for an on-off modulation system with a slicer decision device and an error rate E for both on and off intervals to the peak pulse power P_n required by a polarization modulated system with a differential decision device and also having an error rate E is

$$\frac{P_m}{P_n} = \frac{2K_4U_n\left(1+\sqrt{\frac{2P_B}{K_4U_n}}\right)}{K_1U_n\left(1+\sqrt{1+\frac{4P_B}{K_1U_n}}\right)} \qquad (35)$$

Evaluating the constants, Equation (35) becomes

$$\frac{P_m}{P_n} = \frac{2b_m\left(1+\sqrt{\frac{P_B\eta}{hvb_mU_n}}\right)}{b_n\left(1+\sqrt{1+\frac{2P_B\eta}{hvb_nU_n}}\right)} \qquad (36)$$

This relation reveals one of the peculiarities of optical systems due to square-law detection. For equal bandwidth systems, if $P_B = 0$ (shot noise limited), then $P_m/P_n = 1$ because the slicer decision level W = 0 and the systems are the same. The on-off system is actually better because there are only half as many errors, but this will not change the power required very much. The slicer system also has the advantage that only half as much average power is transmitted. This zero background case is the condition for which the Poisson noise distribution reduces the error rate calculated with the Gaussian approximation so the actual U_N required is about 3 dB less than the assumed value. If the background noise is dominant, Equation (36) gives a ratio of $P_m/P_n = \sqrt{2}$. This is the usual slicer disadvantage of a factor of 2 in the baseband portion of the system, but it is reduced to $\sqrt{2}$ with respect to the received power by the square-law detection. Looking at it the other way the normal 3 dB power gain of baseband polar systems is reduced to 1.5 dB by the square-law detector.

In comparison with the system of Figure D, it has a received average pulse power gain given by Equation (14)

$$\frac{P\text{ (average pulse)}}{P_o\text{ (transmitter output)}} = \frac{2b_m}{J}$$

and a required received power ratio relative to the reference polar system for the same error rate given by Equation (36). The net gain is the ratio of the transmitted pulse powers of the two systems to the ratio of the received pulse powers required for the same error rate.

$$G = \frac{\frac{2b_m}{J}}{\frac{P_m}{P_n}} = \frac{b_n}{J}\left(\frac{1+\sqrt{1+\frac{2P_B\eta}{hvb_nU_n}}}{1+\sqrt{\frac{P_B\eta}{hvb_mU_n}}}\right) \qquad (37)$$

Equation (37) gives the approximate equivalent power gain of the two laser PPM system over a single laser polar system.

Table B lists some calculated values of P_n [Equation (30)], P_m [Equation (34)] and G [Equation (37)] for a specific case. These equations and the calculations in this section involve two major approximations that will require correction factors which are small due to the steep error vs. signal-to-noise ratio curve. The first correction arises because the transmitter pulse power varies depending on the pulse position code from (1+a) to (1-a) where a is 0.177 for the system considered. This will increase the error probability for some pulses and decrease it for others. An

Table B

INFLUENCE OF THE SIGNAL-TO-NOISE RATIO AND BACKGROUND POWER LEVEL ON THE GAIN OF THE TWO LASER PPM SYSTEM

U = 10

P_B	P_n	P_m	G	G(db)
10^{-6}	4.49×10^{-9}	9.36×10^{-8}	195.0	22.9
10^{-8}	4.58×10^{-10}	1.3×10^{-8}	143.0	21.5
10^{-10}	5.58×10^{-11}	4.9×10^{-9}	46.2	16.6
10^{-12}	2.10×10^{-11}	4.1×10^{-9}	20.8	13.2
10^{-14}	2.00×10^{-11}	4.0×10^{-9}	20.3	13.1
		U = 20		
10^{-6}	6.34×10^{-9}	1.35×10^{-7}	191.0	22.8
10^{-8}	6.52×10^{-10}	2.07×10^{-8}	128.0	21.1
10^{-9}	2.21×10^{-10}	1.20×10^{-8}	74.4	18.7
10^{-10}	8.64×10^{-11}	9.27×10^{-9}	37.8	15.8
10^{-11}	4.83×10^{-11}	8.41×10^{-9}	22.7	13.9
10^{-12}	4.09×10^{-11}	8.13×10^{-9}	20.4	13.1
10^{-14}	4.00×10^{-11}	8.00×10^{-9}	20.3	13.1

$\lambda = 10^{-6}$, $\eta = 0.1$, $b_n = 5 \times 10^5$, $b_m = 10^8$

P_B = Watts total received background radiation
P_n = Watts received peak pulse power required for the reference system
P_m = Watts received peak pulse power required for the two laser PPM system
G = Increased transmission path loss for the two laser PPM system

overcorrection reducing the gain by 0.7 dB would maintain the error rate on the weakest pulse. The other correction involves the PPM system. All time slots must be examined for each pulse and an equal probability exists of error or false pulses in all of the 577 slots that do not have pulses. This first calls for a slightly greater slicing level to reduce this probability. This will, for fixed signal power, increase the probability of error for the marked time slot. If the transmitted signal power is simply increased in the case of P_B noise dominant by 1 dB, it will permit the slicing level W to be increased by 52 percent. This will increase the U for the space intervals by 4.6 dB which would be much more than necessary. There are also sophisticated methods of reducing the effect of these space errors since it is known that only one pulse should be recorded. Another point that must be taken into account here is that a pulse interval error causes about 10 bit errors in the H bit stream so that lower error rates must be provided. All these factors should not increase the required transmitter power by more than about 1.5 dB.

For the overall system, the 100 percent conversion of average laser power to pulse power is probably the weakest link. If we allow a correction of 3.5 dB for this, the system gain figures are reduced by a total of 5 dB. This leaves a gain of 18 dB for the high background case and 8 for the shot noise limited case. The gains given in Table B are relative to the power per laser and, since the system of Figure D discussed in detail has two lasers, he actual total transmitted power is two times or 3 dB greater than that of the reference system. This reduces the actual total power gain by another 3 dB. However, the remaining gain is substantial and the multi-laser PPM arrangement appears to be very worthwhile for bit rates in the neighborhood of 10^6 that are of interest for deep space communication. Since b_m has been chosen somewhere near present practical processing limits and the gain falls off with increased pulse rate, it does not look very promising for the bit rates of hundreds of megabits or more required for land communications.

REFERENCES

1. H. F. Wischnia, H. S. Hemstreet, and J. G. Atwood, Determination of Optical Technology Experiments for a Satellite, NASA CR-252, Contract No. Nas 8-11408 with Perkin-Elmer Corporation, July 1965.

2. W. K. Pratt, "Binary Detection in an Optical Polarization Modulation Communication Channel," IEEE Trans. on Communication Technology, October 1966.

3. W. R. Hook, R. H. Dishington, and R. P. Hilberg, "Laser Cavity Dumping Using Time Variable Reflection," Applied Physics Letters, August 1, 1966.

4. A. W. Penney, Jr. and H. A. Heynau, "PTM Single-Pulse Selection From a Mode-Locked Nd^{3+} - Glass Laser Using a Bleachable Dye," Applied Physics Letters, October 1, 1966.

5. S. O. Rice, "Mathematical Analysis of Random Noise," BSTJ, January 1945, p. 130.